Air Quality Data Systems Assessment

(Conducted by Ross & Associates for EPA, OAQPS)

October 2008

ROSS & ASSOCIATES
ENVIRONMENTAL CONSULTING, LTD.
www.ross-assoc.com

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Key findings and recommendations to EPA

- The issue is not that there are too many air systems, but rather that they are operating largely independently, without a common vision. GEOSS provides such a vision at a global level—but it will take specific work (identified in this assessment) at the partner-to-partner level to move towards it.
- As a "user-focused Agency" EPA can provide leadership to this community, by:
    - Investing in a formal partnership with defined goals and shared technical principles. Even a small incremental improvement in science and systems partnering would gain a large return.
    - Making minor re-alignments of EPA systems towards these principles, and, more importantly, establishing them in future investments.
    - Actively support this partnership to help it work. EPA cannot make the needed investments alone, and EPA needs these systems to serve EPA's customers better and more efficiently.
    - A better functioning partnership will help EPA understand its unique role, the roles of other partners and help it focus and target its own resource investments.
- At home, EPA should:
    - Establish an OAQPS Data Coordinator[1]—too much important work is going on without coordination and EPA needs a lead to staff the partnership. This Data Coordinator should also be responsible for improving coordination with ORD (and other science partners) on identifying science priorities and keeping the science-to-operations process working.
    - Refresh the list of OAQPS internal data priorities and establish a systems roadmap towards them. (The Assessment provides specific recommendations on this.)
    - Focus on EPA's fundamental responsibilities where EPA is a data provider: stewardship of the collection process, ensuring easy basic access, and publishing metadata.

Background

In spring 2008, EPA hosted an Air Quality Data Summit of agencies, universities, and consortia to discuss how this community could work more effectively together, towards the common goal of improving Air Quality (AQ) data and its application to environmental decision making. As follow-up to this summit, EPA commissioned this scan of 12 AQ systems to identify how they could operate more effectively together, towards a "preferred future" of inter-operable AQ systems. This report ("the Assessment") provides a summary of the assessment findings.

The assessment developed findings in three areas:

- Sketching a "preferred future for AQ systems"
- Assessing individual systems with respect to the preferred future
- General recommendation to EPA (and its partners)

[1] Development of a standard position description for a role of this type is underway now under the Federal CIO Council.

Preferred AQ data future

We define the AQ preferred future at three levels:

- AQ SoS: It operates as a "System of Systems" (SoS) where the value created and embedded in various resources (e.g., data products and applications) are broadly available via standard interfaces for re-use.
- AQ Partnership: It is supported by institutional partnerships among the AQ community, which helps them make better science and data investments and gives partners sufficient confidence in some common soft and hard infrastructure that they are prepared to use it.
- EPA Mission: It provides EPA with more information about customer needs, and an improved portfolio of investment choices informed by the strengths and capacities of partners.

General findings from system assessment

We developed the following general findings across the set of assessed systems:

- The issue is not that there are too many, overlapping Air Quality Systems, but that they are being developed independently, and in doing so missing opportunities for improvement. When assessed by function, data area, and audience, the systems do not overlap appreciably. We do identify areas of potential future overlap that could be managed with better coordination.
- The basic technical infrastructure for SoS is being implemented already! There are functioning examples of service publishing, mediation and consumption producing new value right now. There are few technical barriers and only a few contentious choices around standards for even broader progress. However, there is limited coordination and most implementations are undertaken piecemeal (i.e., individually).
- End user systems across the board are looking at incorporating both modeled data products and space-based remote sensing data into their applications. This raises both technical and interpretation/application issues and opportunities that are best addressed together. (See recommendations).

General Recommendations

In addition to the EPA-specific recommendations identified above, we recommend the following.

- The ongoing projects and topics that have been informally discussed within the Ad Hoc AQ Steering committee provide exactly the opportunities needed to incrementally move towards the preferred future, but they need to be pursued with more consistency and joint investment.
- Use the metaphor of "AQ Data Value Chains" to identify high-priority flows of AQ data/services, and use these to focus EPA and partner coordination. These value chains will illustrate both where advanced SOA services make sense and also where simple documentation, coordination and flat files will suffice for now.
- EPA should work with partners to develop a shared strategy for prioritizing and promoting better application of *modeled data products and space-based remote sensing data*, and the science-based methods underlying them.

- As an AQ community, we know little about our stakeholders' collective need and data experiences; we need to improve and build new feedback channels for this vital information.

1.0 INTRODUCTION

1.1 Charge

As follow on to the February 2008 Air Quality Data Summit, the Office of Air Quality Planning and Standards requested an assessment portfolio of EPA and related air systems to identify opportunities, inefficiencies, and recommendations as to how EPA and the air quality (AQ) community could take near term steps toward a preferred future. Participants of the Summit identified that while there is ad hoc evolution toward better interoperability from various groups and agencies, there is a need and opportunity to move collectively towards a common future.

This assessment was designed to identify both general and specific recommendations for these systems and EPA, as part of the AQ community, to progress toward a preferred future. It was not designed as a comprehensive inventory of systems and either their scientific usefulness or other domain-specific attributes.

1.2 Definitions

- **Air Quality System of Systems (SoS):** The constellation of AQ-relevant systems which use a System Oriented Architecture (SOA) to interoperate, and which through design and evolution provide improved societal value over the old approach.

- **Tradecraft:** we introduce the term "tradecraft" to identify value-adding functions that data providers or interfaces perform which require special expertise in the relevant data domain, and are therefore difficult for non-experts to reproduce. We recognize that there may not always be agreement on which types of processing merit this distinction. Tradecraft raises special opportunities and issues in the AQ SoS: it makes it essential for high value products to be re-used, and it tempers the perspective it does not matter where processing occurs because it is all generic.

- **Measured Data:** Data taken directly from observations. These data may be aggregated and quality assured or otherwise processed, but the link between the observation and the data is relatively direct.

- **Estimated Data:** Estimated data is based on measured data, but consists of data generated by interpolation, extrapolation and/or by combination with other data based on a given heuristic. Some estimation methodologies are simple; some are very complex and require specific expertise to apply. As these heuristics become more complex, the distinction between modeled and estimated data blurs.

- **Modeled Data:** Modeled data refers to the output of a model run. Model inputs may include measured and estimated (and other modeled) data sources.

1.3 Data Collection & Research

Ross & Associates used the following process to formulate findings and recommendations:

1. Review of assessment questions to generate a Data Collection Guide
2. Initial data collection
    - Review of existing information on systems available on the Earth Science Information Partners (ESIP) website
    - Review of the AQ Data Summit Wiki
    - Interviews with each system's point of contact
    - Trials of the systems (where available)
3. Documentation of initial findings
    - Collection of data in a Systems Comparison Matrix
    - Rescan of information to identify gaps
    - Revised assessment questions (see Assessment Questions)
4. Final analysis
    - Continued collection of data in System Comparison Matrix
    - Design of an Air System Attribute Table
    - Identification of key findings and recommendations
5. Report preparation

Ross & Associates used an iterative approach in gathering data and conducting analysis. This involved comparing the framework of a preferred future with the individual systems. Due to the varied and sometimes ambiguous nature of the original assessment's questions, Ross & Associates revisited and refined the assessment questions based on the data available. These revisions are discussed in the next section.

Preliminary analysis of this data included the production of an individual profile for each system identifying data collection, storage mechanisms, and analysis tool functions.

1.4 Audience

This assessment is written for EPA as an agency and as a partner in the AQ data community; however the majority of the findings and recommendations are applicable to most other partners. A secondary audience for this document is the Ad Hoc Advisory Committee itself.

We assume that readers are already somewhat familiar with these systems and related issues.

1.5 Assessment Questions

Initial research and interviews were based primarily on the *Focused Assessment of Air Quality Community Information Systems DRAFT White Paper,* written by David Mintz following the Air Quality Data Summit. In the White Paper, David Mintz (EPA OAQPS) outlines the goals and expectations of this assessment, and provides ten guiding questions to focus the research. We utilized these questions as a starting point; however mid-way through we reviewed them again to re-assess their value to the assessment, as well as our ability to address them. The questions

where then trimmed to seven primary questions, and supplemented with new questions. These constituted the basis for the *Air System Attribute Table*. The reasoning behind the progression of each original question to the revised versions, as well as where to locate its findings in this document, is outlined below.

Question 1: What are the total annual costs to operate the system? How much is being invested annually in system enhancements, beyond status quo operation?

- Of the system managers interviewed, few were able and/or willing to provide details on their finances; and from those that *were* able to provide estimates, it was difficult to compare results given that there were almost no instances of comparable functions being performed on the same data. It was also challenging to establish consistent system boundaries for the cost estimates provided, and our ability to assess or draw findings from data collected was limited.

- Question 1 was removed from the assessment scope.

Question 2: What would the consequences be if the system became unavailable (permanently or temporarily)?

- This question presumes that there is tight coupling between systems; however the majority of systems (exceptions being GIOVANNI, AIRNow, and DataFed) are *not* tightly coupled. This means that loss of source systems would not immediately impact other interfaces. However, the question brings up discussion of whether system owners actually have a sense of what customers utilize their system for. Simply because a system is used does not mean that the entire business process would stop should it become unavailable. Further, system owners had limited information on users' alternative sources of information. Due to the complexity of the question, focus was placed on the unique value-added of each system and what the implications of *permanent* unavailability would be.

- Findings for Question 2 were built into the system-specific text boxes under *Consequences of System Unavailability.*

Question 3: What factors favor or work against the sustainability of the system (and underlying data) over the next 5 to 10 years?

- Our ability to address this question was based on the findings of other assessment questions, in particular Questions 1 and 5. This data was supplemented by our best judgment and interviews with individual system owners.

- Findings for Question 3 were built into the system-specific text boxes under *Factors Influencing Sustainability of the System*.

Question 4: What function(s) does it serve? Main audience/clients? Regulatory purpose?

› Data on system functions and purposes were available and easily utilized in comparison of systems for potential redundancies. Functions of each system were characterized at a high level, and no notable redundancies were found.

› Findings for Question 4 are addressed in the system-specific attribute tables and the system-specific text boxes under *Core Purpose,* as well as *Value Added to Datasets or Via Interfaces.*

Question 5: Who owns (makes final decisions about) the system, the data, the application?

› Respondent answers were supplemented with our best judgment and information from websites/outside sources. This question is important because it identifies those in the position to influence systems towards the preferred future.

› Findings for Question 5 were built into the system-specific text boxes under *Owners* as well as Recommendation 2.1.

Question 6: How are the data quality assured, particularly the accuracy and completeness of data that is copied or transformed from another data system?

› This issue became more complex than we initially realized, and was not fully covered in the assessment. Examples of problems include systems which copy AQS data using insufficient significant digits, or getting query boundary conditions wrong so that the resulting sets did not match up, as intended with the originals.

› Findings for Question 6 were built into the system-specific text boxes under *QA/QC.*

Question 7: To whom is the data/system available? Are there firewall/security issues that prevent access for certain audiences?

› Most of the systems assessed are public (un-authenticated public). The focus of Question 7 in this assessment was placed on where there were significant data sensitivity issues which would motivate security layering beyond basic user self-registration.

› Findings for Question 7 are found in the System-Specific Findings on EIS and RSIG.

Question 8: What data standards are used, and how do those coordinate or conflict with standards in the other data systems, particularly the systems that provide primary data intake and archiving?

› The critical aspect of this question is determining if there are fundamental incompatibilities between systems, with missing/different metadata, missing/different data, or incompatible code sets or data concepts. Standards in use were identified at a high level; however data

standards themselves were not raised as a key limiting issue by any system owner. Many systems make use of multiple overlapping standards, and for many areas it is not clear there is an "authoritative" standard or that their use would resolve any fundamental semantic incompatibilities.

› Question 8 was removed from the assessment scope.

Question 9: How easily can the system be adapted to a GEO or Exchange Network interoperability framework?

› GEOSS and the Exchange Network were interpreted as flavors of the same overall architecture: SOA. GEOSS advocates a general service oriented architecture with emphasis on Open Geospatial Consortium (OGC) web services; so to a small extent, any OGC, Web Coverage Service (WCS), or Web Mapping Service (WMS) is GEOSS "compliant." The EN has similar architecture with more extensive specification of the web services themselves. Because all systems assessed store data in Relational Database Management Systems (RDBs), the real question is: could services be made accessible as external web services of other flavors?

› Question 9 was removed from the assessment scope, but is, in effect, addressed by the overall assessment.

Question 10: Does the system copy and store data from another source?

› A study of the systems found that, in many cases, "copy" was an oversimplification. Focus was placed on the QA/QC of the copy process as well as the value added to copied data.

› Findings for Question 10 can be located in the system-specific attribute tables and contributed to findings on Primary Data Providers.

1.6 Caveats

Assessment of system performance and user satisfaction

A major qualification for our assessment is that we did not attempt to judge the quality of system performance or assess user satisfaction with these systems. If the systems interface claims some functionality, or was identified by the system manager, we assumed it does that function reasonably well and that users are reasonably satisfied. As discussed in the findings, many of these systems have relatively limited ability to obtain much information about users or their satisfaction.

Restricted portfolio of systems

The Air Community of Interest (COI) is composed of many more systems than were included in this assessment. Our assessments about "unique" value-added and other recommendations apply only to the portfolio assessed; however we think that many of the general

recommendations and findings can be applied to any system operating within this domain. It may be beneficial to consider a future assessment that conducts a systematic review of all systems within the Air COI, beyond the scope that was possible here.

System Attributes Scan Proved Subjective

This assessment was not designed as an inventory to support discovery or description. The assessment questions focused on less tangible aspects such as management or relative emphasis of functions. The comparisons provided are not determinations of "adequacy" of any system. Early reviewers all comment that they want more information in this table (for discovery and description), we make a recommendation that this be pursued as a community, in the near term.

1.7 General Findings

1. A central charge of this assessment was to identify overlap between the systems examined and to recommend ways of eliminating duplicated functionality. We found no significant overlap among system functionality; each system provided a unique set of features that were specifically tailored for a different user group. Rather than overlap, this assessment did identify existing and emerging gaps in the data and functionality offered by the systems. Access to model runs and remote sensing data is incomplete and often impractical. User and web service interfaces to the data were more often incomplete than duplicative, indicating that there is still significant opportunities for development work in the SoS. We see this as a major opportunity for EPA and the Air COI generally, and have tailored recommendation 2.1 to address this.

2. Generally, the systems examined can be grouped into data providers and data consumers. In some sense, every system included in the assessment provided data, either to other systems or to users; however there were four systems whose sole purpose was to be the primary repository of record for regulatory data. These systems were functionally different than the data consuming systems, with the majority of the infrastructure and resources devoted to a large database. Generally, these systems were not easily accessible by the public, though the data was publically available. The remaining systems we examined were data consumers—i.e., their primary purpose was to provide data products to a specific group of end users. These systems often included unique data; however it was usually derived from the measured data provided by the primary source systems (VIEWS is a notable exception).

3. All systems are considering, if not implementing, web-services for data consumption or data publishing. This indicates that systems are moving closer to an ad-hoc SoS; however there is not an over arching architecture plan to help guide the systems to a *preferred* future.

1.8 Sketching Attributes of a Preferred Future

Sketching attributes of a preferred future for AQ data community

As discussed in *Data Collection & Research*, this assessment attempted to simultaneously refine a framework of a preferred future for AQ data while comparing existing systems individually to that future. Most discussion (before, during, and after the AQ Data Summit) on the AQ community around a preferred future has focused on the "pumps and pipes" aspect of the basic concept of SoS. Under this framework, AQ data "services" flow from provider applications to consumer applications; these applications then in turn do something useful for their users. To the discussion of a high level architecture, we have little to add and do not include an SOA introduction here.[2]

The assessment seeks to broaden the preferred future to include additional issues such as the relationship between the "IT" side of an AQ SoS and the science underpinning AQ data products. It also seeks to provide EPA with specific options for next steps on its own individual systems.

Because participation in any future AQ SoS will be voluntary for many actors, we focus on the choices potential participants, specifically EPA, might make. As illustrated in Figure 1, below, choices are considered in three domains:

1. The choices EPA makes internally on its information system investments, programs, and policies. Recommendations 2.3 and 2.5 focus on this domain.

2. EPA's choices and investments in coordination/partnering with its AQ partners around all topics, including coordination of the development and operation of an AQ SoS. This includes joint planning to identify the highest merit value-chains for AQ data and users. Recommendations 2.1 and 2.2, as well as Chapter 3 address this domain.

3. The AQ SoS itself, its standards, protocols, and architecture. This is where individual system assessments as well as the majority of conversations have focused in this assessment. This topic is developed throughout the report, specifically in Recommendation 2.4 and the System-Specific Findings and Recommendations.

We identify these three domains explicitly to highlight specific real world investments and policy decisions the governmental partners could, or should, make to support the preferred future. Within each domain, goals relating to technical infrastructure and data product creation are equally important, and we attempt to highlight the necessary approaches and methods to realize those goals. Eventually partners will have to argue to their internal governance and/or funders as to why this approach makes business sense for their individual system and specific

[2] Note that our use of the terminology SOA or SoS does NOT pre-suppose either the specific transport or payload standards used. We observe this community converging on OGC standards and a small set of payload standards (on which more work is recommended). The organizational and basic workflow coordination for these systems will determine success of the larger SoS, not the adoption of this or that technology or standard.

mission. Ultimately, the sum of these decisions will determine the success or failure of the SoS. This issue is discussed further in the Domain 2 discussion below.

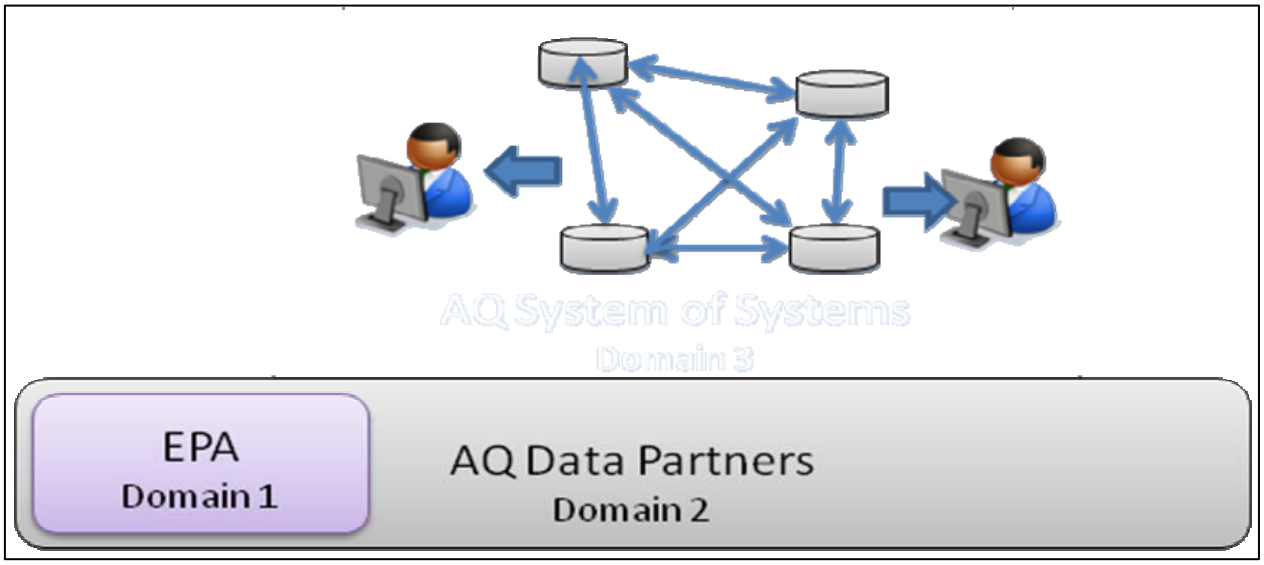

Figure 1: Domains for a Preferred Future

Sketching a preferred future framework for each domain

Domain 1: EPA Mission

Attributes of a preferred future for EPA:

› EPA priority AQ customers (internal and external) can find, understand, and use AQ (EPA and other) data to achieve societal benefit. Doing this will require:
  - Better use of the data to provide relevant and responsive information to decision makers.
  - Increased ability of decision makers to apply the new information generated from the SoS to achieve the societal benefit.

› EPA benefits more from existing investments and is able to make better investment choices because it has:
  - Better investment options, which play to strengths of partners, avoid redundancy, and achieve synergy. (See Recommendation 2.1)
  - Better information regarding customer needs, levels of performance in existing SoS, and new data sources and systems. (See Recommendation 2.1)

› EPA and EPA-funded data assets, analytic processes, and other resources are, to the greatest extent possible, available to authorized internal and external users in a format/venue useful to them.

› EPA has sufficient confidence in core SoS soft and hard infrastructure to rely on it for some functions.

› A sustainable process is put into place to transition EPA research data products and services into well-described operational services on the SoS.

› EPA can effectively communicate the societal benefit of the components of the existing SoS, demonstrate impacts of its loss, and articulate a vision for its future development.

Domain 2: EPA and the AQ Data Community

This domain includes how broader policy level coordination between EPA and its other federal partners provide societal value, specifically around how the large, long-term investments in science and monitoring. In addition to the preferred future attributes listed below, which focus on organizational support of the SoS, this domain also includes broader issues such as coordination on science objectives and desired products, and participation in the even broader world of GEOSS. Most interviewees indicated these issues as the "upstream cause" of some of the inconsistencies we identified in the system assessments. While not a focus of this systems-based assessment, these issues were raised often enough that we included attributes which address them. Success here determines what new sources of data or whole systems will exist in the future, or which aspects of the current systems are sustained.

Attributes of a preferred future for EPA as a partner in the AQ Data Community:

› All major investments made by EPA and partners in air quality information systems will produce new data and/or data services which are available over a set of standard interfaces to authorized uses, in addition to whatever dedicated applications are developed.

› Through joint planning with other partners and stakeholders, investments are targeted on key value-adding functions/priorities. (See Recommendation 2.2)

› User requirements, usage rates, and user satisfaction information is collected and made broadly available to inform partner decisions.

› Partners have confidence in the functionality of core SoS soft and hard infrastructure. This confidence will lead to partners integrating public functionality into new applications and, in turn, making new applications available to the broader community. This will also create an environment in which partners pursue funding opportunities that rely on the SoS infrastructure. Ultimately, the confidence in the SoS will encourage partners to co-invest in the development and maintenance of shared systems with EPA.

› Partners are coordinating to provide incentives (e.g., via granting and recognition processes) to users of the SoS.

› A sustainable process is put into place to transition research data products and services into well-described operational services on the SoS. This allows the community to broadly utilize specialized data knowledge and skills (tradecraft).

Domain 3: AQ System of Systems

Attributes of a preferred future for the SoS itself:

- Operates as a dynamic constellation of loosely coupled, SOA connected systems.
- All services are entered into a common discovery service which is available to any relevant part of the SoS.
- All data sources make their data, at all relevant stages of processing, available via standard web service interfaces in addition to whatever interface applications are developed.
    - Data transformations/fusions or integrations, especially those which are creating new, value-added data products, will be available via a standard web services interface.
- SoS architecture and access services are agnostic about data storage locations of their clients. Both ad hoc query and local-copy-refresh oriented services are available where appropriate.
- Consumption of many public services requires little to no coordination with the provider.
    - Metadata needed to bind to and interpret the data are readily available.
- Data and Interfaces are based on open standards, beginning with but not limited to OGC/GEOSS, and evolving as those standards do (or should[3]).
- There is widespread re-use of information and services, and little avoidable duplication of effort in system and data processing development.
- Core components of the SoS are formally supported and transparently operated so that partners trust their sustainability and are prepared to rely on them. Core components include:
    - Discovery service
    - Shared security service (if any)
    - And mediation services for managing interface incompatibility/version lags
- SoS operation still allows agencies/partners to establish/retain key socio-political returns such as attribution, credit, publicity, or reputation in having their data products and services flow through the SoS rather than going it alone.

[3] Existing WCS 1.0 standard alone, is not meeting needs of many users to provide finer grained query capability, this is an active area of WCS development.

Finding and Recommendations for These Domains

The table below provides a crosswalk of these domains and the recommendations developed.

Table 1: Preferred Future Domains and Recommendations

Domain	Recommendation	Description
1 EPA	2.3: Roles of EPA as a Primary Data Provider	Recommendation 2.3 outlines a role of EPA as a data provider providing stewardship of the collection and value-adding processes to both the data and metadata, making data resources available to users in a useful format as described in Domain 1.
1 EPA	2.5: EPA Systems: Development Paths, Internal Architecture, and Strategy	Recommendation 2.5 identifies improved coordination opportunities for OAQPS architecture.
2 AQ Partners	2.1: Evolve an AQ Data Community to Improve Shared Planning.	Recommendation 2.1 identifies collaborative efforts in the AQ data community which lead to the prefered SoS as descirbed in Domain 2.
2 AQ Partners	2.2: Identify and Support Key Value Added Products in the AQ Data Value Chain	Recommendation 2.2 focuses on a set of specific value chains from which a coordinated SoS could be based, as detailed in Domain 2.
2 AQ Partners	Chapter 3: AQ current projects are important places to work (or work harder) on these issues	Chapter 3 highlights immediate opportunities in current projects to advance this domain.
3 SoS	2.4: Increasing Distribution and Use of Modeled Data	Recommendation 2.4 outlines availability of modeled and other gridded data, as Domain 3 discusses through standard web service interfaces and SoS architecture.
3 System Findings	Appendix A: System-Specific Findings and Recommendations	Appendix A provides system-specific findings and recommendations towards the SoS.

2.0 GENERAL RECOMMENDATIONS

2.1 Evolve an AQ Data Community to Improve Shared Planning and Implementation Capacity

Air Summit participants identified a "…need [for] some group constituted with enough decision making authority to establish stability and confidence in the standards and protocols on which the system would rely." Our interviews and findings for each system reinforce the need behind this goal, specifically because:

- There is general acceptance of the common vision of shared services. Systems are already modifying existing, or mounting new web serve interfaces to provide additional access. However this piecemeal evolution is occurring without cross-partner coordination. This presents a salient opportunity for the community to grow value of the current data products and services by collective planning and development of a shared system design.
    - While there is a clear convergence on the Web Coverage Service (WCS) and Web Map Service (WMS) specifications, within each there is the still potential for significant variation in the way like data types are defined and accessed. This increases the cost of using these systems, limiting wide scale adoption. An example is lack of interoperability between the WCS interfaces used by RSIG and the WCS interface implemented by DataFed. These services, as currently defined, do not implement a common definition of a coverage, leaving potential users to abstract the different conceptual models behind each in order to combine the data. We recognize that improvements in technology make "brokering" between different services simpler; however, for most air system users the benefit to having an agreed-upon set of standards for publishing WCS and WMS in the short term is clear.

- System managers told us that interagency prioritization of data needs is a continuing opportunity for alignment and coordination of these systems. Within this domain, there is no other overarching coordinating group that can look across institutional boundaries for the purposes of shared planning. This presents a significant opportunity to leverage funding and institutional resources across agencies towards a common future. EPA, NASA, and NOAA are the three primary data producers for this COI, and the data produced by each is clearly complimentary. Additional planning and identification of key data needs could dramatically improve the coverage and quality of data sources, expanding both the breadth and depth of data available to the COI. This is a key component to realizing any common vision for a preferred future.

Proposed agenda for the ad hoc AQ steering group (See Chapter 3)

Based on these general findings, we recommend that the ad hoc steering group should continue as the central coordinating body of the Air COI. This group is roughly representative of the major systems; and flexibility as well as informality are critical in the very near term. We provide a list of issues and a straw agenda for this group in Chapter 3. Highlights of this include:

- Finish the process initiated at the Air Data Summit. Draft a shared aspirations/expectations document which includes a vision of the preferred future, and identify key roles and responsibilities within that preferred future. This report outlines some of those attributes, based on our analysis of the different systems.

- As identified in the preferred future section, there are obvious SoS functions that are needed, such as data discovery and description. Most of this assessment has focused on access, but these other functions are just as important. The preferred future section provides initial starting points for shared implementation work around each of these functions, which should be examined by the coordinating group.

- There are likely scores of web service implementations underway or being planned now that could be improved with coordination. The responsibility of the Ad Hoc group would be to take a "systems" perspective on these implementations and provide advice to implementers.

- Establish a "Science to Operational Data" standing topic area for discussion of relevant activities and plans, which ad hoc group is already doing informally. This would be a good connection point for the specific GEOSS prototypes discussed in Recommendation 2.2.

Suggestions for gathering user feedback

Participants at the AQ summit ranked the topics of improving the collection and sharing of user requirements and feedback in the top tier of shared priorities. With some exceptions, the systems assessed here capture very little user information. At the surface, this is a challenge for individual systems, however a community approach to developing user feedback methodologies is recommended. Improving information on customer requirements and satisfaction is hard enough for one system, let alone the many systems participating in the AQ SoS. This is a challenging, "soft" problem that will need a solution from within the COI. The goal of these methodologies should be to consistently measure user feedback on features and services in a common way that can be reused across systems.

Several starting points are:

- For areas that involve many partners, encourage partners to open/share their requirements development process/documentation via a common collaboration space. This information could be linked in a common collaborative space to the systems inventory process started here.

- Consider adoption of some mechanism by which users (or their machines) are encouraged to register so that usage patterns can be tracked and periodic solicitation for feedback and recommendations on new features/fixes can be emailed. VIEWS provides a good example of how this can be implemented. Private sector services typically require such a self-authenticated key, validated only by an operational email, for access. This, at the very least, provides contact information for feedback.

- For many services, a wiki/bulletin board serves as the primary feedback mechanism; this approach could be standardized, at least to the extent of the discovery service providing a link to that site.

› For internal EPA systems, we recommend EPA implement a structure for collecting and analyzing user feedback. We recognize that a major challenge to implementing a system is the lack of a past mechanism for collecting this information. We see two main components of this system:
  - Identifying users of the web, Exchange Network, and SQLNet services of a system. Web logs and EN NAAS account information may already collect this information; however it needs to be centrally aggregated to be useful.
  - Gathering usage patterns and feedback from users.

We acknowledge that user surveys are rarely filled out and thus are not usually worth the effort of their development and distribution. Instead, we recommend that EPA pursue a proactive method for gathering user feedback. One potential avenue for data collection is the Air COI national conferences and data summits. These venues provide an excellent opportunity to collect information and feedback on how users interact with a system and their suggestions for improvements/modifications.

EPA and partners should consider placing more emphasis on the requirement of transparent user tracking (with opt out) in the systems they fund. It's likely that a small increase in attention to improving the quality of this information would yield results. This pressure will likely have to come from outside, since system developers will often prioritize feature expansion over an administrative support like this.

2.2 Identify and Support Key Value-Added Products in the AQ Data Value Chain

What is the AQ data value chain?

This recommendation steps back to ask what data and flows are most important. It is based on the assumption that a relatively small number of particularly important AQ data objectives and associated data products/services could be identified by the AQ community, without a need for agreement on their relative value. As they do for some current agency and cross-agency efforts, these products/services could provide as useful focal points for broader joint planning of the AQ SoS. Delivery in whatever their form of these products should be high priorities for the AQ community.

In preparation for the AQ summit, a basic "value chain" diagram was developed for the AQ community (shown below). This diagram was an adaptation of several diagrams, including those depicting DataFed and VIEWS.

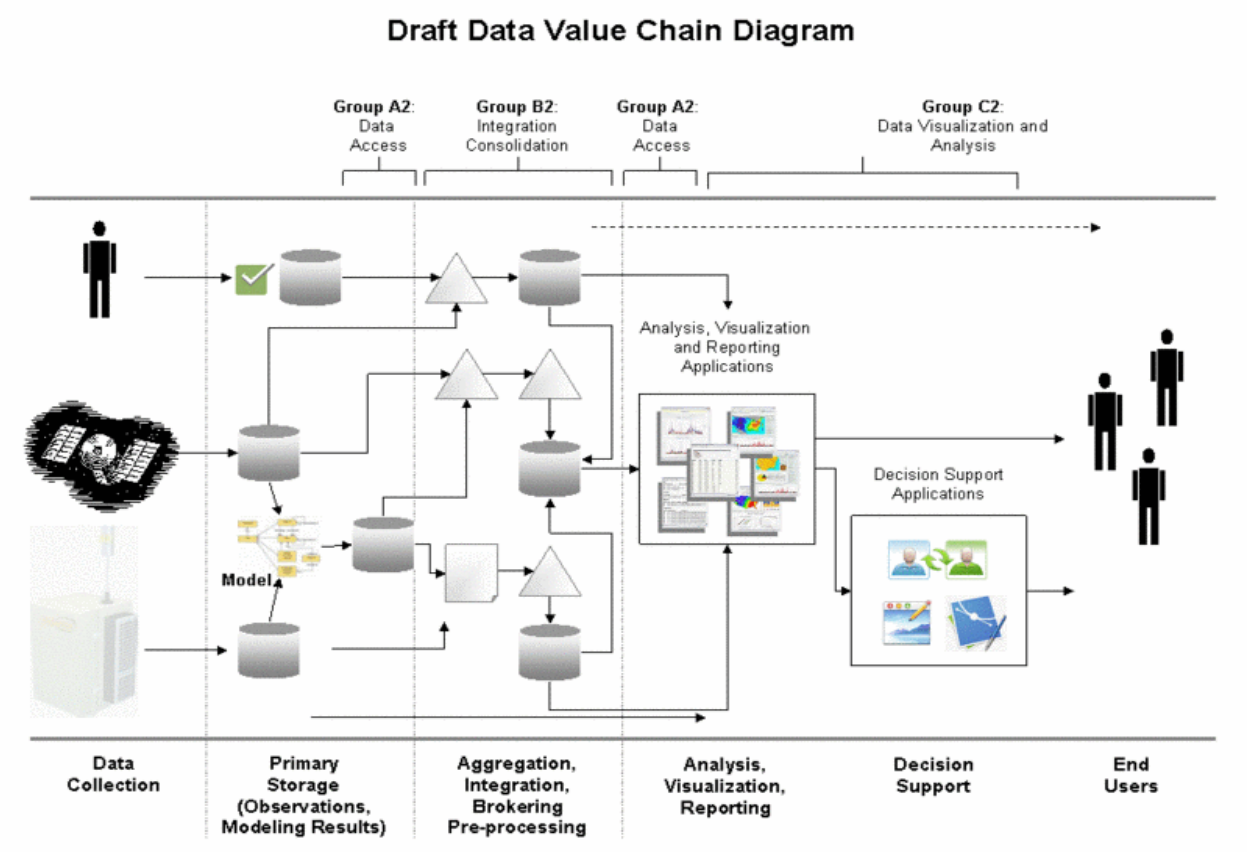

Figure 2: Draft Data Value Chain Diagram

This diagram was used to organize conversation around the basic concept of a value chain for information, and around a set of common high-level Information management functions that are routinely performed at various stages (such as data collection, aggregation, etc.). The

diagram was also intended to motivate discussion of a "loosely coupled" architecture for the preferred future where functions could be performed on a distributed basis, with many different interfaces, using many common services—depending on their relative strengths and weaknesses.

As far as it goes, this generic diagram is useful; however it fails to depict much about ***how*** these functions add up to societal value via the final product. These generic IM functions alone failed to capture much of the real value-added of the systems assessed. We need a simple theory of how "value adding" works so we can provide some non-obvious insights into how design of an AQ SoS can maximize it.

For example, knowing that a given system does integration and fusion of data says very little. On the other hand, knowing that a given system is the sole place where a given, respected, heuristic for such fusion is implemented, and that the resulting data product is provided on a regular basis, says a lot. On the basis of the limited portfolio of systems assessed, we identified a set of straw intermediate AQ information functions which seems to cover many common data objectives. These functions, and how they were served by various systems and initiatives, followed a common pattern:

- Development of or improvement of the basic underlying science/technique used to collect/analyze the data (e.g., research on remote sensing estimation techniques or on dynamics underlying a model).

- Applying and/or improving the application of this science/technique to a specific problem domain (e.g., remote sensing to surface parameter estimation, interpolation of ground level measurements to produce AQ surfaces, method inter-comparison to develop cross-walks).

- Using that science/technique to produce or apply a product for an end result (like an analysis or support of a decision) in that domain.

Figure 3 was developed to summarize these main functions observed in the systems examined:

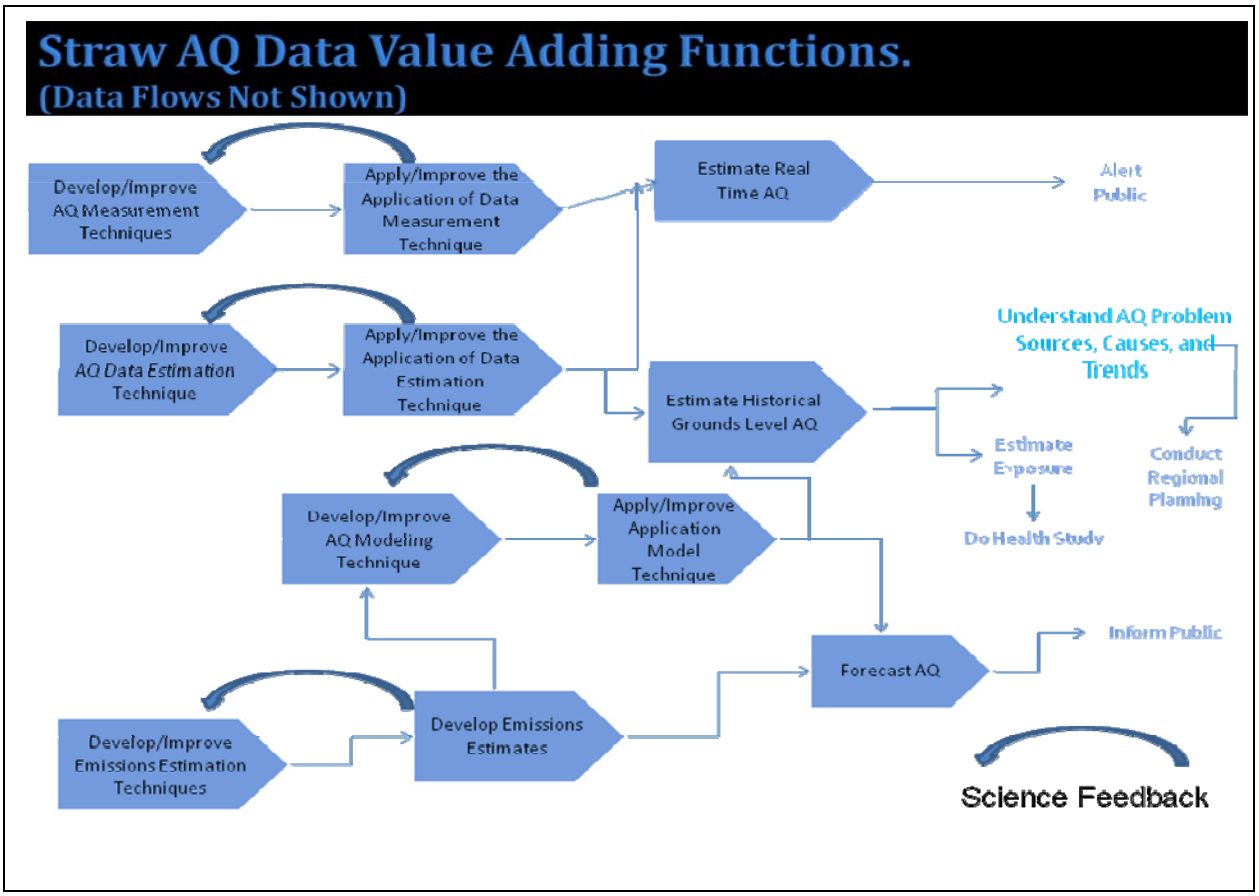

Figure 3: Straw AQ Data Value Adding Functions

The Table below provides some examples of how some specific systems support these functions.

Table 2: System Function Support

System Function Examples	System(s)	How Supported
Apply/Improve the Application of Data Estimation Technique	AQS	AQS stores multiple method estimates for key parameters to allow inter-comparison.
Estimate Real Time AQ	Air Now	Aggregates and re-distributes Real Time AQ Data Next steps: improving interpolation for real time provision of AQ surfaces.
Develop Emissions Estimates Develop/Improve Emissions Estimation Techniques	EIS	EIS will store multiple method estimates for emissions estimates allowing for comparison, and improvement. Next steps: Air Quality Decision Support System (DSS)[4] project to improve use of RS to provide area emissions estimates

Example: Providing data to the health community

As described in the System Findings for the AQS system, this approach was motivated by tracing the value chain for the familiar scenario of health researchers struggling to obtain air quality data from AQS to estimate historical surface concentrations and, in-turn, estimate exposure. What that researcher really wants is a well documented data source/service for these historical estimated concentration surfaces using one or more known estimation techniques.

Given this context we can ask the AQ SoS the hypothetical question: *Who is responsible (each year) for ensuring that the best known techniques for combining remote sensing data, with the best techniques for surface data estimations are packed as a common general purpose product for the relevant practical scales?*[5] While some progress has been made on providing the health community access to parts of such a product, we think it deserves more direct consideration as a priority (if the basic scenario is accurate) especially because it is likely a commonly desired product type of product across all Air domains.

We suggest this only as an illustrative example; however it may not be the most relevant or important.

A focus on the "middle tier" of these value chains: science of application

In the diagram above, the "middle tier" deserves special focus:

- As illustrated by initiatives such as the DSS project and applications like RSIG, a key common theme is operationalizing research data/techniques, specifically advanced modeling and remote sensing parameter estimates. This is one of the most important things new applications do. Where they are successful, the SoS provides the means to distribute those products widely. These projects, and this focus, represents the opportunity for organizational learning by the AQ community as a whole, in identifying both the soft (methods) and hard products (applications/interfaces) that support this objective.

- The feedback arrows highlight the imperative of informing science priorities via operations experience. During the AQ summit, the depiction used of this "AQ community intelligence" was of feedback arrows going from right to left in the value chain, informing up-stream components of requirements and lessons learned.

Summary recommendations

- OAQPS should (with its partners) prototype and validate this approach with a small set of specific value chains to identify priorities. This was attempted at the AQ data summit but with mixed success, perhaps in part because the linkage between the result and the preferred future were both unclear. Hopefully this assessment and ensuing AQ community discussions have clarified this connection, and readied the community to try again.

- Ability to link individual system (system components) to these value chains could be a powerful tool for discovery and description.

› This approach would also be a powerful complement to a "gap" analysis that could be conducted from an inventory of:
    - Value chains supported
    - Spatial temporal coverage
    - Parameters
    - Basic functions (Discovery, Description, Integration etc.)

› The ad hoc group could scan current "portfolio" of shared projects/initiatives to identify the value chains they implement and better understand how they can be supported with services (see Chapter 3).

2.3 Where EPA has a Primary Data Provider Role, it should Focus on Some Basics

We identify the following high-level roles of EPA as a primary data provider:

1. Stewardship of collection process and system.

2. Stewardship of the value-adding processes run by EPA that produce new data products/interfaces; stewardship of the metadata/documentation that describe all data products and interfaces.

3. Ensuring that key user groups have the access they need (to raw data, value-added data products, and other services).

4. Providing unique knowledge/assistance on data use that cannot be provided by others.

In support of these roles we identify two key areas that should shape the AQ SoS:

Primary Source Data Publishing: As the owner of several large data systems that the air community depends upon (AQS, AIRNow, EIS), EPA is in a unique position to improve the services and formats upon which the air community moves towards a preferred future. Therefore, we recommend EPA focus, above all else, on strengthening its access services for the "gold standard" primary source data. As discussed in Recommendation 2.3, this should include:

› Implementing the next generation of AQ web services

› Commissioning development of a general-purpose measured AQ data navigation, query and download tool, using these services[6]

EPA data are widely republished and so new services would create significant benefit to a large portion of the air community. Additionally, building out the current SOA to implement WCS and WMS services would require only a small development effort. This is an opportunity for EPA to take the leadership in implementing the preferred future, by soliciting community input on design and implementing rapidly. Additionally, these services offer an opportunity for EPA to consider how best to publish value-added data products developed under Recommendation 2.5. These products will present a new set of challenges for the agency regarding scope and volume of data, and thus the development of a new generation of services will be well positioned to aid the agency in considering solutions.

Implicit in this recommendation is the de-emphasis, for EPA, of developing COI-focused user interfaces. Our system-specific findings as a whole suggest that domain-specific systems and organizations are better positioned to understand and meet the needs of a specific community. Creating robust publishing services will make building specific interfaces simpler, and, we find, would be a better allocation of resources, given that smaller organizations are not able to provide the amount of primary source data that EPA does.

Metadata:

It seems there is never enough metadata and what there is, is the wrong kind. Despite this dilemma, metadata is still one of the most important supports for effective discovery and use of data. Recognizing this, EPA and the Air community generally have attempted, and struggled, to create methods to voluntarily collect and publish metadata. For EPA systems, this metadata has already been partially collected; information compiled by EPA staff represents a rich source of

data. However, there are still significant gaps that need to be addressed. The system assessment has given us the following findings regarding the difficulties of metadata:

- There are limits to purely "voluntary" community efforts to collect metadata. Despite its great successes, DataFed's "Wiki" approach to metadata, relying on the community as a whole to contribute metadata about the Datafed sources, has not worked very well.

- Operational metadata: in addition to better exposing the method-focused metadata (where it exits) the assessment also identified an equally important type of "operational" metadata that needs improvement.
  - EPA should, as soon as possible, establish, document and publish its procedures for updating ALL of its primary data feeds, even (especially) if they are just flat files. Much of the uncertainty around some of data in DataFed stems from neither side knowing the when and how of these data feeds.

- Additional metadata observations:
  - Given the diversity of user expectations we recommend that AQS focus on the most basic and generic metadata, with an emphasis on supporting discovery and basic description.
  - Publish that metadata as a web service to support its integration into other applications.
  - Build in a suggestions solicitation for enhancements to that metadata.
  - Include "operations" metadata, especially about the update frequency and mechanisms. External systems owners we interviewed highlighted their interest in this information.

EPA should also sponsor collaboration on a community-wide approach to metadata management and publication. Of all the systems we examined, we did not find one solution to metadata publishing that had achieved widespread use. This is an area where EPA should focus development effort. Because there is no model from which to build, we recommend that for the first generation of metadata publishing services, EPA should focus on a system that addresses 80% of the need, rather than attempting to create a perfect system that addresses 100%. The relative benefit of implementing a system that meets most user needs greatly outweighs the additional costs of meeting every potential user need.

2.4 Increasing Distribution and Use of Large Estimated and Modeled Data Sets

This finding codifies and emphasizes a trend we think is clear to most AQ Data observers—that there is going to be increasing appetite and use of modeled data, and other gridded data products in the near future, and that these products need to be considered simply as additional forms of "data." These types of products present known issues and opportunities in how they are produced, distributed, and interpreted. The discussion below identifies some common themes identified across the systems for this issue. This recommendation covers two very different data types together: large datasets as in space-based remote sensing, and model outputs. While fundamentally different in origin, they have many issues in common about how they are managed, distributed and interpreted.

Growth Observed and Projected

For many of the applicable systems assessed, integrating or enhancing the use of modeled/estimated data products was identified as a near term priority. These included:

- VIEWS : incorporating both additional remote sensing products
- RSIG: including CMAQ model outputs
- CASTNET: enhancing modeling to produce more robust estimated deposition surfaces
- NASA: Research Opportunities in Space and Earth Sciences (ROSES)
- AirNow: (pending funding from NASA) integration of remote sensing products
- EPA/CDC efforts to establish common AQ concentration products for use by health researchers

We submit that because this area is poised for rapid advances, and because many of the issues identified below have been partially characterized and are familiar with the community, that they represent a good place for early joint planning and infrastructure coordination, an opportunity to get "ahead of the curve" as a community and avoid an unnecessarily piecemeal approach.

Looking a bit farther out, are the "3D" model and measurement data. These sets present even larger technical and interpretive challenges. Programs such as the EPA-GEO Advance Monitoring Initiative and the NASA Applied Science Program seek to prototype the "middle tier" function discussed in Recommendation 2.2, of testing real world applications of these new data sets for improved decision making. There is still debate about the relative merit of investing in these types of data, vs. improvements in delivery and application of existing "2D" data. A role of the Ad Hoc group could be to ensure that these disagreements to not impede the SoS itself.

Key issues

Technical/Transport Challenges

Modeled data (and other estimated data) are typically large-gridded data products which cannot be effectively served up whole, and whose total size often precludes local storage of all

data of possible interest. These requirements match up very well with the basic SOA assumption of the preferred future: the key is providing access to this data through granular services, as both RSIG and GIOVANNI have done. Additionally, the need for these services presents an opportunity for the community to collectively consider how the existing services could be enhanced in the future (See Recommendation 2.1).

User Sophistication/Access and Model/Data Interpretation

Historically (and currently for EPA), only the people who run a model can access its raw output. Less sophisticated users are never exposed to model outputs unless they are highly distilled/summarized into another product. The technology itself has allowed the community to avoid some more sensitive issues of disagreements about the validity of a given model or run, or should be working with models and their output more directly because the technology barrier moderated access. This situation is changing. For example, RSIG currently hosts *Community Multiscale Air Quality* (CMAQ) model products on the development side of RSIG (both data access and visualization). RSIG is weighing if this will require adding a simple user authentication process to limit access to these resources once moved to the operational side of RSIG. This situation is somewhat similar to existing user screening doing by systems for partially screening some data.

Metadata Required to Support Interpretation/Integration

There are a very broad range of possible metadata objectives one could have with model products. Of specific import, in the near term, for many of the systems assessed, was the development of sufficient (and ideally somewhat comparable) metadata to be used for two very specific purposes:

- Support their operational integration into new interfaces such as RSIG, VIEWS or other.
- Support high-level descriptions of these products to the low-/medium-sophistication users accessing these products over these interfaces.

As discussed at the AQ Data Summit there is much activity, but only very limited coordination on the development of metadata conventions to support these needs. This coordination is recommended as a near term action area, which, per Chapter 3, can be executed in the course of existing projects.

Summary of recommendation

OAQPS should develop a strategy for how to host and ensure access to modeled (gridded) data

- EPA should supplement its own experience with these data types with the extensive experience of its partners. Per the proposed next steps in Recommendation 2.4, EPA should identify some near-term priority products (probably the new health-researcher targeted PM concentration surfaces) and pilot new approaches to providing access. Initially EPA may elect to simply make a set of large files available for FTP to known partners who can serve that data up via standard interfaces using WCS.

- Per Recommendation 2.5 OAQPS should include evaluation of how it can increase access to CMAS model outputs internally.

- EPA should carefully track the existing portfolio of projects which are dealing with precisely this issue, for practices which can be expanded.

Primary barriers to broader use of modeled data output may be political/institutional

- Anecdotal information suggests that many barriers to broader model output publishing are not purely technical.

- AQ model out puts are often "tightly bound" to the systems and institutions that produce them. EPA has limitations on what it can publish in various capacities, and many modeling organizations may be resistant to release un-interpreted model outputs for facile re-consumption.

- EPA should place these issues on the agenda of the AQ community for input. Some solutions will require new conventions, but there may be technological supports such as binding metadata, attribution, and recommended use data using a standard approach this may make more participants comfortable with broader distribution of "preliminary" or "research" grade results.

2.5 Recommended EPA Systems Roadmap

This section focuses on summary findings and recommendations specifically for OAQPS (and related OA) systems. Many of the other recommendations (e.g., increasing use of modeled data) also apply to OAQPS but are discussed in their respective sections.

OAQPS needs to renew a roadmap for its systems

EPA requested that this assessment include EPA systems in order to:

- Identify opportunities for improved coordination and effectiveness internally.
- Inform, align with and support the emerging AQ SoS.

Over the past 15 years, OAQPS staff and management identified the growing need for access to AQ and supporting data via an integrated interface. This common need motivated some increased attention to coordination across the office and the development of AirQuest as a general purpose central integrated repository and set of business analytic applications. The original vision of AirQuest was quite broad. Much of this planning and coordination function has, as of recently, lapsed. More recently, EPA launched the AQS DataMart, the EIS system development process, and improvements to AIRNow; however these developments have been conducted mostly independent of one another. While some project-based interface development for AirQuest has continued, development of a general high performance interface to it has paused.

Recommendation 2.1 outlined what we see as EPA's primary responsibilities for providing access to its core data programs (AQS, AIRNow, CASTNET, and EIS). This recommendation focuses on aspects of EPA's internal architecture and coordination of these systems to achieve these goals, and to provide EPA with internal capacity to make better use of these data.

Within this context we identified the following findings and recommendations:

1. **Establish an OAQPS Coordinator and use it to staff data coordination within OAQPS,** also including ORD and other relevant OAR offices. This position would also be lead staff for support of the AQ partnership.

2. **Re-affirm original vision of AirQuest:** The original objective of AirQuest—to provide OAQPS users across the office with integrated access to these data sources, and to use this infrastructure as a complement to efforts to better coordinate these investments—is sound and relevant. Demands on EPA analysts are, if anything, increasing.

3. **Advances in access technology mean many external resources are now available as services:** As EPA staff have observed, the rapid developments in data integration and exchange technologies occurring externally, including emergence of applications like GIOVANNI and DataFed, mean that there are many external resources EPA can tap for internal customers. These resources may not need to be warehoused at EPA.

4. **Refresh the characterization of EPA OAQPS analyst needs and future program decision needs**. OAQPS should refresh its characterization of OAQPS analyst needs, in sufficient detail to inform the target architecture. This characterization should also include a look at likely future program needs. Some of these needs will be unique, but many will overlap with those of external customers. OAQPS will have to make a judgment as to the internal executive support for the investment required to implement this restructuring. Some common challenges include:
   - Sourcing and integrating remote sensing data into local applications/interfaces, is AirQuest still the right solution?
   - Managing common access to Modeled data (and other gridded products (see Recommendation 2.4)
   - Using best available techniques to process/integrate data sets once for re-use.
   - Stewarding better metadata data to support better understanding by secondary internal and external users.

5. **Framework for OAPQS Architecture:** The data integration needs of OAQPS will be satisfied by a combination of warehousing and application integration; the important question is which systems do which. We recommend the following as a starting point:
   - Unless there have been significant flaws identified to date, use the AQS DataMart as the integration point for AQS, AirNow and CASTNET data. (We recommend renewing efforts to get CASTNET into AQS so that EPA offers a unified source, and so that access application/metadata investments made by EPA can be leveraged. Perhaps EPA can even offer the CASTNET users a new version of their abandoned dynamic interface at no incremental cost).
   - Evaluate AirQuest's embedded integration business logic and decide which of it is best kept embedded in the current platform and which could simply be migrated. The AQS data in AirQuest already comes from the DataMart, and the NEI data, in the next generation should come from a new NEI warehouse. It is a distraction to compare or critique the current AirQuest interface, because it has received no recent investment, and is not where the hard work has been done.
   - Per the above, plan immediately how/where to warehouse the NEI data from EIS. We recognize that this is not on EIS's near term priority list but it needs to be on OAQPS's.
   - From the analyst needs analysis conducted above, identify an incremental pilot that can be conducted for providing some common infrastructure for model product distribution—we did not review any system relevant to this finding, but identify it as a component of the needed internal architecture, OAQPS needs to know enough about what and how of these data to know what hooks need to be implemented for integration with the sources identified above.

A sketch of this transition is provided in Figure 4.

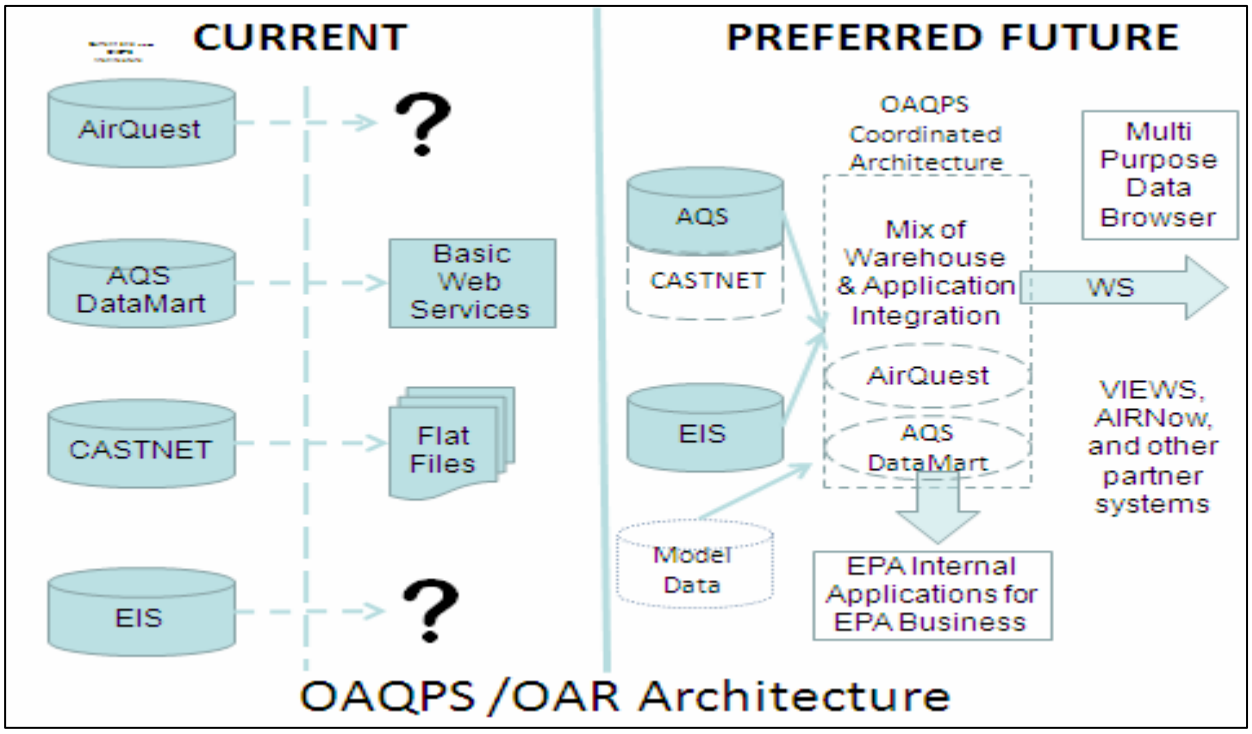

Figure 4: OAQPS/OAR Architecture

6. **Build capacity to compare/contrast data estimation and model products:** As illustrated in the re-designed EIS, a critical capacity for OAQPS will be the management and comparison of multiple data products using a variety of data estimation and modeling techniques so that analysts can understand their strengths and weaknesses. Given the increasing diversity and availability of methods, OAQPS needs to improve its capability to manage this diversity not just select from it.

7. **Consider this need has an analog in EPA work with OAQPS external partners:** ideally OAQPS can play a role in incrementing towards a preferred future where productive comparisons of various techniques can be conducted, evaluated, and translated into new data products and services. It is in OAQPS's broader interest to see that this "market place" is effective.

8. **OAQPS should communicate assessment results to EPA's OEI in order to:**
   - Identify where current EPA policies and architecture are barriers to better access/integration.
   - Identify services/support OEI could offer, especially in connection to the Exchange Network and expanding use of web services, to support OAR internal and external customers.
   - Illustrate the dire need for an alternative architecture to support high performance, semi-public, service interfaces to EPA data. OEI has identified and discussed possible services and architecture changes for nearly all the issues identified above. This case study would be helpful to that effort.

3.0 CURRENT OPPORTUNITIES TO CONSIDER AND IMPLEMENT ASSESSMENT RECOMMENDATIONS

There are many projects underway in the Air COI, and likely many more that we are not aware of, which provide an immediate opportunity for EPA and its partners to act together towards these recommendations. This section highlights some of these opportunities

Community Components for the Air Quality Scenario in the GEO Architecture Implementation Pilot – Phase 2 (AIP-2)

- This project proposes service catalog, portal, and metadata system. All areas identified in the original Data Summit and re-confirmed in this assessment. The AQ community needs a non-ad-hoc system for data description and discovery across sources. A starting point could be to syndicate and aggregate existing system-level machine readable metadata from partners. EPA has spent significant resources developing the Registry of EPA Applications and Databases (READ) that could be used to provide the EPA feed for this approach.

VIEW Enhancements:

Key aspects of current VIEW enhancements include:

- The VIEWS user community and the systems feedback communication from this community could provide a useful test bed (for this user type) for evaluating the utility of various analytic and visualization approaches.
- The broader issue of integration of remote sensing data, within a decision support application is being addressed now, and should prove instructive to the broader community.
- VIEWS is considering now the what/how of exposing more system resources via services. Most users of VIEWS either use the application itself, or download data manually for local analysis. VIEWS (and the community) would benefit from a broader conversation.

Air Quality Decision Support System (DSS) [NASA/ROSES funded project)

This project was discussed at the AQ Data Summit and the Ad Hoc group, it advances many of the key value chains identified in Recommendation 2.2, including:

Assessment Areas	Project Objective
Improving emissions estimation process. Routine integration of remote sensing data	1. Routine capture, analysis, and processing algorithms with high temporal and spatial resolution to provide land use/land cover data as inputs to emissions and air quality modeling analyses.
	2. Acquisition of satellite data to obtain increased temporal and spatial resolution of activity data and emission rates from natural and anthropogenic area and point sources (both individual and clustered) in remote and urban areas.
Recommendation 2.4	3. Incorporation of three- to four-dimensional (2-3 spatial, 1 temporal) pollutant fields (e.g., aerosol extinction profiles, column NO_2 and O_3), into the DSS to improve boundary

	inputs and evaluation of outputs from gridded chemistry-transport models (CTMs) such as CMAQ.
Recommendation 2.4	4. Development of advanced analysis tools for examining the satellite data to better understand the relevant atmospheric processes and their representation in the CTMs.
"Middle Tier" in value chains, focusing on application within a given decision domain.	5. Visualization and quantitative analysis of satellite data in combination with existing monitoring and emissions data, and modeling results within a unified data analysis and decision support platform.

As indicated in this table, this project will provide many opportunities for the creation of new data resources/services, as well as new application. The critical role of the AQ SoS would be to encourage and support the conscious design of these services for broad re-use, from the beginning. This discussion regarding service based provision of the remote sensing data is especially relevant to GIOVANNI and RSIG.

AQS DataMart/AirNow Data File Format Coordination

- AirNow and the DataMart are exploring ways to identify a common, improved file format for transmissions between the two systems. This could be used as service basic FTP downloads and as the payload for web services mechanisms. Flat file submissions will continue to be a viable method of transferring large datasets, and the DataMart and AirNow are seeking to use a common format for exchanging datasets. This presents an opportunity for the entire community to learn from this experience and possible provide input on other success with large file formats.

Straw Near-term Agenda for the Ad Hoc Advisory Group

We propose that the Ad Hoc advisory group focus on the goal of hosting a new Data Summit in the summer of 2009. Preparations should focus on:

1. Formalizing and affirming the Ad Hoc group's focus and scope.
2. Developing a common vision for the preferred future.
3. Joint implementation of the near term opportunities discussed in this assessment.

Below is a straw for agenda items leading up to a new 2009 Data Summit, we submit that a pre-condition to this summit is some tangible progress on a critical mass of these areas. Central to each of these items is a commitment of staff time to conducting the background research needed to inform the discussions. Without appropriate preparation and background, the Ad Hoc group will be unable to move forward.

[Note: we have NOT reviewed the pending EPA solicitation for related coordination services, this proposal and that would obviously have to be considered together]

Action Item	Description	Time Period
Develop statement of purpose formalizing ad-hoc steering group as formal governance body for the air community	This should be the first product of the Ad Hoc group, outlining the goals for the next year and major products that need to be created. Proposed purpose of the group is to provide a forum for information sharing and development of best practices.	Q1
Finish shared aspirations/expectations document	This document was started after the AQ data summit. The Ad Hoc committee should use this as a first step to formalizing the Air COI around the Ad Hoc group and affirming the statement of purpose.	Q1
Create a first generation list of existing WCS/WMS, and establish an aspirational "to be" list of such services	The Ad Hoc group can serve as the forum for sharing experiences implementing web services using the WCS standard. Existing work now stalled in the interop group should be re-invigorated to coordinate existing implementations.	Q1
Refine system wide preferred future for discussion at AQ Data Summit 09	Using the ideas presented in this report and other sources, the Ad Hoc/Interoperability Sub Group should articulate and affirm a SoS wide vision of the preferred future.	Q2
Establish a proposal for first generation shared discovery and description infrastructure.	Provide a framework for data discovery and description (and its relationship to GEOSS) that can, at the minimum be used as the point of departure for upcoming system investments/	Q2
Establish a "Science to Operational Data" standing topic area for discussion of relevant activities and plans	The Ad Hoc group has already begun this step informally. This would be a good connection point for the specific GEOSS prototypes discussed in Recommendation 2.2.	Q3
Jointly identify (but not prioritize[7]) list of key value adding AQ data functions. **Use these to organize 09 AQ Data Summit**	The Ad Hoc group should identify the common key value adding functions that are done across many systems. Following this report's Recommendation 2.2, the value adding data functions may then be expanded and republished as services.	Q3

APPENDIX A: AIR SYSTEM ATTRIBUTE TABLE

The goal of the Air System Attribute Table is to provide a high-level overview of the features and attributes of each system. We developed a set of 13 categories based on the original 10 questions and additional information obtained through interviews.

This table is designed to complement the system-specific findings that follow in the next section, though some information is characterized only in the Air System Attribute Table. For easy reference, we have included each system's Attribute Table at the top of the relevant system findings section. Below each Attribute category, we have provided the rationale for the individual rankings.

The Air System Attribute Table categories broadly cover system architecture, design, function, governance, and usage patterns. It is not meant to be a complete list of the attributes of each system. As described in the Caveats, there are some areas that we were not able to gather information on, like relative user value/satisfaction.

Lastly, as identified in the Caveats above, the values were based on our own judgments derived from interview notes and other background research conducted. We used only three categories in these rankings to reflect our estimated precision of the assessment.

The scores provided for the system attributes are meant to be informative, not evaluative. They contain a large measure of best professional judgment—for example, some of the scores are inherently relative to system size, and a large system scored at ½ for a given attribute may in fact have greater "magnitude" of that attribute than a small system scored at full.

Air System Attribute Table Category Discussion

- **Data Update Frequency:** The length of the regular cycle in which the data is updated or added to the data system. Systems may update components more frequently than this on an ad-hoc basis

- **Unique/Primary Data Provider for that Assemblage of Data:** This column reflects a categorization of the major functionality of the system. We examined four systems whose sole purpose was to be the repository of record for measured regulatory data. Other systems included some unique data, however this was generally a secondary function. Note that VIEWS is considered a primary repository for data from the IMPROVE network and this data is in AQS as well. A full circle in this category indicates that the sole purpose of the system is to be the primary repository of a particular data set (example AQS). A half circle indicates that the system includes a unique data set, but providing access to that data is not necessarily the sole purpose of the system. An empty circle indicates that substantially all that the data in the systems is available elsewhere.

- **Created as a "One Stop" for data from multiple systems:** Those systems that were not designed solely as data repositories were generally created as a "one-stop" solution for a particular community of interest. Some systems have evolved from this design philosophy to focus on a particular type of data or value-adding functionality. A full circle indicates that this system was architected as a 'one-stop' location for a subset of data gathered from multiple systems.

- **System Includes "Novice" User-Friendly Integrated User Interface:** This is a purely subjective judgment of the usability of the primary interface for a novice user. This may not be a relevant class of user for some systems.. Some systems did not include user interfaces, and others were designed exclusively to be a user interface to republished data. A full circle indicates that the system includes a very user-friendly interface that is accessible to all potential users. A half circle indicates that the user interface is functional, but may be too complicated for the 'Novice' user. An empty circle indicates that the system doesn't have an end user interface, or includes a very complicated user interfaced designed for experts in the data domain.

- **System Includes Outbound Web Services (used by external customers);** this is a general metric regarding the availability of any web service interface to the system. A web service is considered to include any SOAP-based service, any OGC service, Exchange Network services, and HTTP (REST) services. A full circle in this category indicates that the system provides some sort of Web Service interface to the data. An empty circle means the system does not provide a web service interface.

- **Formal Group Governance/Process:** This is a subjective measure of the governance process tied to a system. We considered group governance to be any formally-recognized group of stakeholders who had a significant role in providing direction and management of the system. A full circle indicates that the system has a group governance structure independent of the funding/owner agency composed of stakeholders with decision making authority over the direction and priorities of the system. A half circle indicates an a more informal governance structure in place, however it may be only advisory. An empty circle indicates the system has no standing governance process in place.

- **Formal Feature Development Process Including Community Input:** Related to the category above, this metric was an attempt to identify systems that systematically collect and make use of user feedback for identifying and prioritizing features and architecture. All systems did this informally to some degree; however we see value in a formal approach to gathering user feedback. Generally, those systems with some sort of formal governance group also had a formal process for gathering input. A full circle indicates that the system implements a formal design process that includes content from stakeholders and the user community. A half circle indicates that the system gathers feature requests and development ideas from users on an ad-hoc basis. An empty circle indicates standing system for feedback from user community.

- **Public Access to System:** This category indicates whether the system provides public access to the interface/data. We considered systems that require registration only partially public, even though anyone may be able to get access upon registration. Some systems, particularly regulatory primary sources systems, did not allow general public access, for varying reasons. We found most systems provided at least some level of public access. A full circle in this category indicates that a system provides complete access to all features and data to the public without registration. A half circle indicates that the system provides access to all features and data to registered users, and access to a limited set of features/data to unregistered public users. An empty circle indicates that the system does not allow access to public users.

- **Behind EPA Firewall:** This category simply indicates whether the system is currently contained behind the EPA firewall. A half circle in this category indicates that a component part of the system may be behind the EPA firewall, even though the interface is not.

- **Different User Access Levels:** This category is related to "Public Access" above. Most systems that limit public access to the system do so through an authorization scheme involving different levels of user access. A full circle indicates that the system implements granular permissions and access levels for different types of users. A half circle indicates that there is only one level of access for registered users. An empty circle indicates that there is no system for regulating user access based on authorization.

- **Integrated Interface Used for AVR:** AVR stands for Analysis, Visualization, and Reporting. This is a standard processing chain for data systems. A full circle indicates that the system includes an interface with full AVR functionality. A system represented by a half circle may only implement some AVR functionality. An empty circle indicates that a system does not include an interface, or that the interface includes no AVR functionality.

Table 3: Air System Attributes

	Data Update/ Generation Frequency	Unique/ Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process	Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
AIRNow	Hourly	●	○	●	●	◒	◒	●	○	◒	●
AIRNow Tech	Daily	◒	○	●	●	◒	◒	●	○	◒	●
AirQuest	Varies	○	●	○	○	○	○	◒	●	○	◒
AQS	Quarterly	●	○	○	○	◒	◒	○	●	●	○
DataMart	Daily	○	○	○		○	○	◒	●	○	○
CASTNET	Quarterly	●	○	○	○	○	○	◒	◒	○	○
EIS	Annual	●	○	○	○	◒	◒	○	●	●	(In Dev)
RSIG	Varies	◒	●	●	●	◒	○	◒	◒	◒	●
VIEWS	Varies	◒	●	●	○	●	●	●	○	●	●
DataFed	Varies	○	●	◒	●	○	○	●	○	○	◒
GIOVANNI	Varies	○	●	●	●	●	●	●	○	○→◒	●
HEI	Every 6 months	○	●	●	○	◒	●	●	○	○	○

APPENDIX B: SYSTEM-SPECIFIC FINDINGS

System-Specific Findings: AIRNow

AIRNow Attribute Table

Data Update/ Generation Frequency	Unique/ Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Hourly	●	◒	●	●	◒
	Primary storage for real time version of Ozone and PM2.5	Centralizes and aggregates forecast data from 125 agencies hourly, however data is from a single domain	Access is provided through AirNow Gateway	Outbound web services including WSDL/SOAP through the AirNow Gateway. Currently developing next generation of interface	Steering Committee input from EPA, regional offices, state and local agencies

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
◒	●	○	◒	●
Decisions made at EPA, keeping stakeholder input in mind	Registration is required for access. Public domain EPA website. Respects the wishes of the data provider as to whether particular datasets can go public or not.	Hosted outside by Sonoma Tech	Some data providers do not want public access to their data; these datasets only available internally	Outbound provided through AirNow Gateway (the only web app) and text files. Also provides an RSS feed and raw data. Working on other outbound services, like AirNow International.

AIRNow Tech Attribute Table

Data Update/ Generation Frequency	Unique/ Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
daily	◒	◒	●	●	◒
	Sole provider of national forecast information	Created to provide web-based access to AirNow	Access is provided through AirNow Gateway	Outbound web services including WSDL/SOAP through the AirNow Gateway; currently developing next generation of interface	Steering Committee input from EPA, regional offices, state and local agencies

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
◒	●	○	◒	●
Decisions made at EPA, keeping stakeholder input in mind	Registration is required for access. Public domain EPA website. Respects the wishes of the data provider as to whether particular datasets can go public or not.		Some data providers do not want public access to their data; these datasets only available internally	Outbound provided through AirNow Gateway (the only web app) and text files. Also provides an RSS feed and raw data. Working on other outbound services, like AirNow International.

AIRNow & AIRNow Tech Profile

Core Purpose	Serves as data collector, provides access to real-time ozone and criteria pollutants data. Aggregates, processes, and distributes AQ data from 125 agencies hourly.
Value Added to Datasets or via Interfaces	Near real-time AQ measured data and estimated AQ surfaces. Centralizes and aggregates processes and data. AIRNow Tech performs some integration.
Factors Influencing Sustainability of the System	Voluntary system supported by EPA but hosted externally. Super high-visibility system. Used to support AQI, embedded into media outlets, and used extensive as primary source for real-time AQ data. Sustainability of AIRNow Tech is linked to AIRNow.
Owners	EPA designed and operated system
QA/QC	Utilizes a quality control system
Consequences of System Unavailability	Serious consequences for system unavailability. Given AIRNow's tight temporal coupling to all major news/media and governmental notification/alerting systems, there would be an immediate impact.

Findings and Recommendations

› The program/system commonly referred to as "AirNow" is composed of three parts:
  - AIRNow Website system for distribution of real time AQ data and images. This site, via various approaches, is widely used by the public, media, and NGOs to source AQ data.
  - AIRNow Tech password-protected site focused on supporting and aggregating of AQ forecasts
  - AirNow Gateway subsystem of FTP, HTTP file access and web services which serve up AIRNow data.

For expedience we've combined these in the assessment. However most of the observations about improved integration with the AQ SoS relate mostly to the AirNow and the Gateway.

› AirNow is the go-to system for real-time AQ data. Its reputation and success at forging partnerships with providers, and relatively stable funding, indicates it is highly sustainable. We see it as the real-time analog of AQS as a primary source data system.

› AirNow is hosted at a contractor facility, outside the EPA firewall. This approach is unusual, but it alleviates many of the access/interface challenges of other EPA systems.

› For the assessment, we consider AirNow in two roles:
  - As outlined above, as a primary data provider for real time AQ data
  - As a COI support application for AQ monitoring and forecasting staff around the country

› In this second role, AirNow continues to develop analysis, integration and display tools designed to support interpretation and use of the hosted data. This currently includes aggregation of met data, historical AQ data from AQS, and pending funding (see below) remote sensing data. While targeted at information complementary to real-time data analyses, these features will naturally begin to overlap with other systems providing AQ AVR functionality. We identify this as an opportunity for re-use of services developed either by others or AirNow for these functions, or, if that does not happen, as a possible redundancy.

› AIRNow's experience with Web Service interfaces may provide a useful lesson for the broader community. For some time AirNow has had a rich, outbound SOAP-based Web Service[8], but it has had few users—more users probably use the DataFed connection to AirNow. In its current incarnation the service is structured as a single large (rather complex) service. AirNow staff are re-factoring this service into smaller, granular services which should be easier to invoke. Staff also intend to distribute pre-built simple client applications (e.g., Excel spreadsheet macros, java scripts) that can be easily incorporated into desktop or web applications.

Recommendations Summary

› AirNow has a reputation for effective and rapid development, but also for operating somewhat independently. Given the reputation and flexible development environment available, AirNow should be leveraged by EPA as key place to aggressively mount and integrate services. Details are suggested below.

› As of October 2008 the ad hoc's interoperability group discussion on mounting of WCS is somewhat stalled, it should be re-invigorated and used to develop and demonstrate WCS and WMS flows from AirNow. AirNow is working on this now but doing so, out of necessity, independently.
  - This will immediately provide RSIG and others with an improved access mechanism. AirNow staff should participate with AQS staff and then mount a WCS to be used for AQS DataMart population and consumption. Getting this data directly from the DataMart would eliminate a duplicate and potentially conflicting store of historical AQS data.

› AirNow adoption by the international/open source community represents a critical long term opportunity for the entire community. The architecture and tool set for this release is under development now.
  - As planned now, international use of AirNow will use a "hub and spoke" model, which, combined with the data standards built into the system, means there will improved access to real time data for many new sources.
  - Development of an open-source community potentially provides lower-cost distribution mechanism for tools, and new people power for development.
  - All these will depend upon/could leverage improved articulations of priorities from secondary data consumers.

› AirNow has a funding proposal in to NASA to consume and integrate images and data. This illustrates that as this type of data/images becomes more available, it will proliferate in various interfaces. This will require technology solutions similar to that demonstrated in RSIG: rapid, possibly distributed data sub-setting and display.

System-Specific Findings: AirQuest

AirQuest Attribute Table

Data Update/ Generation Frequency	Unique/ Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Varies	○	●	○	○	○
	Secondary only	Complex value added integration			

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
○	◒	●	○	◒
Management team makes decisions for system development. No longer building applications; inviting analysts to look at the database and build applications to go on top of it	Some external access provided to the NASA-funded 3D AQS project and University of Maryland Baltimore County			Created an online map that users can access from Desktop ARK GIS Software, custom developed application

AirQuest Profile

Core Purpose	Provide integrated access to emissions, AQS data and a wide range of supporting data, in order to establish a common information base for OAPQS analysts and their applications.
Value Added to Datasets or via Interfaces	Key value-added in AirQuest is the business logic used to support complex geospatial queries of NEI data, integrated with AQS data. Only unified source within EPA of NEI/AQS. Knowledge embedded integration of these datasets.
Factors Influencing Sustainability of the System	Internal EPA system, so sustainability is determined by OAQPS. System has fewer internal supporters.
Owners	OAQPS: Management structure with monthly meetings. Three management co-leads, one senior management advisor, and members of the team.
QA/QC	Not assessed.
Consequences of System Unavailability	Internal EPA impacts.

Findings and Recommendations

See Recommendation 2.5 for additional discussion of AirQuest.

AirQuest was developed in response to:

- A long-range OAQPS planning effort to identify and support internal analyst data needs.

› An interest in bringing some consistency to analysis (being done on separate desktop systems) by developing a common information base and applications.

The goal of changing the culture of analysts working mostly separately with separate information bases was acknowledged to take more time and more intervention than just making a new system available.

The current AirQuest system combines a wide range of data sources, including AQS, NEI, AIRNow, REMSAD model runs, and GIS layers, as well as census and meteorological datasets; it dimensionally integrates the data spatially and temporally for rapid query access. Through the NASA funded 3-D AQS project, AirQuest data sources include MODIS, GASP, and MISR aerosol optical depth data.

AirQuest is now primarily in maintenance mode with only limited, special-purpose interface development underway. The update processes for the system are mostly automated and annual operating costs are low.

The goal of widespread adoption of the system by analysts has not been achieved; however the system is used by several interfaces/application with OAQPS. As indicated elsewhere, we believe that the key value-added aspect of AirQuest is its integration of business logic and its capture of a historically-relevant set of OAQPS analysts' needs, including the data locator and metadata navigation tools. Below we recommend that AirQuest and its use be re-examined in the larger context of OAQPS systems, to determine why it is not being used. Possible reasons for non-use include:

› That the assumptions driving the creation of AirQuest are correct, but people are not using it, (in part) because of the lack of functional interface.

› That data in AirQuest is lacking in some way—e.g., lack of resolution, undocumented/trusted integration, or update techniques. The design of AirQuest may have underestimated the need for more fine-resolution data. It stores most data at the county level.

› The data in AirQuest is incomplete—some primary source data is missing, though available elsewhere.

Some of the hard work has been done developing approaches to deal with the need to consider arbitrary geometries like attainment areas. One option to consider is merging that portion of AirQuest with the DataMart. This would allow access to finer-grained AQS data already dimensionally modeled in the DataMart.

Recommendations Summary

› We recommend that the future of AirQuest be considered in the context of other OAQPS systems. The fundamental question is whether the AirQuest approach the right mechanism for providing integrated access to the identified data sources, given the new services available.

› We identify the following relationships to other system trajectories:

  - AQS DataMart: AirQuest copies its data from the DataMart but could also be re-factored to simply use the DataMart. This may address a perceived weakness of the system lacking sufficiently finely-grained AQS data.

  - NEI/EIS: AirQuest's use of EIS data may be informative in the design of an EIS warehouse. If such a warehouse is developed, AirQuest could be re-factored to use it.

  - AirQuest was identified under the NASA 3-D AQS project as the system to house relevant satellite remote sensing data sets. Currently AirQuest contains coincident satellite pixel data match to ground monitors but there are additional large data sets that have been provided to EPA but have not been put into AirQuest.

System-Specific Findings: AQS and AQS DataMart

AQS Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
quarterly	●	○	○	○	◒
	Primary source for official ambient AQ measurements	Created as a computer repository for ambient air quality measurements from thousands of ground-based monitoring stations	Provides flat file and web page of text files		EPA operated, internal AQS project team

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
◒	○	●	●	○
Developments are largely based on what is heard from talking with users	Only EPA employees can have SQL Net access		Access is primarily through DataMart using the EN. Also provides flat files	

AQS DataMart Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
daily	○	○	○	●	○
		Main purpose is access to AQS, also contains AIRNow data		A web service interface that uses the EN version 1.1 standard, netCDF, and CDF	

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
○	◒	●	○	○
Developments are largely based on what is heard from talking with users	Access through Exchange Network Node. External users may also request web service access to the Data Mart. EPA employees and on-site contractors may request SQL*Net access. Access is available to the public, but system is not geared toward that.			

AQS Profile

Core Purpose	National database of ambient air quality measurements.
Value Added to Datasets or via Interfaces	Sole source for EPA official ambient air quality data.
Factors Influencing Sustainability of the System	Core EPA regulatory support system.
Owners	EPA designed and operated system.
QA/QC	Documented QA/QC process on inbound submissions.
Consequences of System Unavailability	AQS data feeds into DataFed, AirQuest, RSIG, VIEWS, HEI, and AIRNow would stop.

AQS DataMart Profile

Core Purpose	Primary access to AQS data as well as AirNow data. Integrates AQS and AirNow data.
Value Added to Datasets or via Interfaces	Provides complex aggregation/low integration and some decision support to users.
Factors Influencing Sustainability of the System	Currently has limited users/supporters but is also currently the *only* such AQ access point (alternative is flat files or direct dump from AQS). Sustainability controlled by OAQPS.
Owners	EPA designed and operated system.
QA/QC	Documented QA/QC process on inbound submissions.
Consequences of System Unavailability	AQS Data feeds into DataFed, AirQuest, RSIG, VIEWS, HEI, and AirNow would stop.

Findings and Recommendations

AQS

AQS is the repository of record for national ambient air quality measurements. The AQS application is designed primarily for management of the data submission process and for moderate-/expert-level access to the data via query and reporting routines. The focus of this assessment has been on access to AQS data. While many users still access the AQS system for data, we anticipate (and encourage) that migration of all data access functions *not* related to the data workflow process be sourced from the DataMart. The assessment team did not look at the access tools in the AQS application.

AQS data is perhaps the most reused data of any system examined within this assessment—it is republished through no less than seven other systems. For this reason, the AQS system is one of the most sustainable systems examined, and conversely would negatively affect the Air COI if it became unavailable.

DataMart

The DataMart should be a major provider of publicly-available AQS data. It offers Exchange Network (SOAP/WSDL + EN) outbound data services, as well as SQLnet access to trusted partners. The DataMart does not yet, however, publish data through web coverage services or other OGC standard interfaces, the emerging de facto transmission mechanism among other air community partners. Given the original service-oriented design of the DataMart infrastructure, adding additional interfaces should be low cost but create extremely high value for the Air COI as a whole.

The DataMart EN services currently offer data in XML format. Staff dedicates some time developing custom datasets in netCDF, CDF, and other formats. The fact that a broad range of users were utilizing HEI system's downloads of AQS data (see HEI System-Specific Findings), combined with the fact that EPA staff still spend significant time developing CDF and other flat files for users, indicates that there is an appetite for a simple, user friendly interface to access simple low-/medium-end access to this data. Collecting information on the necessary data formats and providing them as output options would increase the value of the DataMart services to end users.

As discussed in Recommendation 2.2, many consumers of AQS raw data intend to use that data to estimate area concentrations; for example asking for the zip code of monitor locations (not stored in AQS). In some cases it would make sense to make those users aware of existing estimated data products (e.g., those being jointly developed by EPA/CDC and others) which might better suit their needs.

In many cases, EPA is dependent upon others for the metadata on individual measurements in AQS. EPA has compiled significant amounts of metadata about the data in the DataMart; however, no easily-accessible mechanism for publishing this information exists. Given the significant amount of republishing occurring and the volume of users, this is a major shortcoming which impacts the way this data is used.

› Publishing the AQS metadata that has been collected to date would be a major value-adding product to the Air COI. Given the number of systems that consume AQS data from the DataMart, metadata publishing services could have a significant audience. Existing metadata could be enriched in two ways: first, by improving the accuracy and volume of self-reported metadata, and second, by publishing metadata as a service available to all DataMart users.

Recommendations Summary

› The AQS DataMart needs a general access user interface that provides query and dataset download to users in a common format.

› Per Recommendation 2.3, EPA should work to improve the outbound data publishing services of the DataMart. Additionally, the AQS metadata that currently exists should be easily accessible through a publishing mechanism to enable better use, and reuse, of DataMart products.

System-Specific Findings: CASTNET

CASTNET Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Quarterly	●	○	○	○	○
	Primary source of alternative air quality information.	Created to get a more accurate picture of ambient air quality			

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
○	◒	◒	○	○
	Access through public FTP site; does not require registration	EPA firewall via SQL*Net and ODBC drivers. Data are available via the Web as prepackaged data sets and a data query wizard.	No	Used to provide an interface, but found it more cost effective to go back to providing flat files

CASTNET Profile

Core Purpose	Collect and publish current and historical dry deposition data from approximately 80 backcountry location sites located around the country. .
Value Added to Datasets or via Interfaces	Data is combined with meteorology data collected on site and summary aggregates are calculated by contractors before submission to the database. Current estimated interpolations are supplemented with more advanced CMAQ-modeled data.
Factors Influencing Sustainability of the System	Core system for EPA mandated program. Primary source of alternative air quality information. Used by CAMD and multiple other sources to compare AQS data and supplement/ground truth model results.
Owners	EPA Clean Air Markets Division.
QA/QC	Contractor performed QA/QC on dry deposition monitors on through the reporting process.
Consequences of System Unavailability	Lack of regional picture of air quality. VIEWS and DataFed no longer have access to CASTNET data.

Findings and Recommendations

CASTNET is a primary source data system that provides 20 years of historical dry deposition data for the entire country. Key data uniqueness is the location of the monitoring sites: all are placed in remote areas, including national parks and forest lands, providing an accurate picture of regional ambient air quality uninfluenced by local sources.

CASTNET is a partnership between EPA Clean Air Market Division, USDA Forest Service, DOI National Parks Service and Bureau of Land Management. The system is comprised of the monitoring sites and a database system. Contractors gather the dry deposition data weekly

from the monitors, analyze the results, and send the data to CAMD for collection and storage. The entire dataset of current and historical data is available through a public FTP site.

The data from CASTNET is primarily used by researchers as a comparison/validation of AQS data and model results. Additionally, CASTNET data is used to provide high-level surface maps for the continental United States. There is significant value for CASTNET datasets as a new source of regulatory/enforcement data and primary input for model runs. CAMD is working to use CMAQ model run results to fill out the available CASTNET datasets.

Prior to this year, CASTNET data was available through a dynamic web query interface online. This site was developed and hosted by EPA OEI. Due to significant problems with data synchronization and service responsiveness, CAMD opted to abandon the web interface in favor a public FTP site hosting the entire dataset available as flat files, for an estimated savings of 40k per year. Given the target audience (mainly researchers interested in the entire dataset) it is unclear as whether this choice has had any negative impact on users' ability to access and retrieve data.

Recommendations Summary

- There is clearly an opportunity to improve access to CASTNET data. If plans for integrating CMAQ model run results with the monitoring station data move forward, the volume of data available in CASTNET will rapidly make FTP access to the entire modeled dataset impractical. In keeping with the SOA component of the preferred future, a medium-term priority should be creating web service interfaces to CASTNET because the dataset will be too large to access in one piece; granular queries will be the only feasible method of accessing data. The simplest solution will be to provide granular access to the data through a WWW or web service interface.
  - OAQPS should work with CAMD to scope the potential for integrating CASTNET data into the DataMart.

- Making the CASTNET data more easily available for republishing would add value to the Air Community generally, as the CASTNET data is the second major source for this type of data. As modeling increases generally as a source of new data for the community, the use of CASTNET data to verify model results will increase.

System-Specific Findings: EIS

EIS Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
annual	●	○	○	○	◒
	EPA's primary compilation and estimation of emissions	Some additional data included but focus in on emissions	Planned		Primarily EPA directed, EIS Task Force, input provided from PQA and the States

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
◒	○	●	●	(in dev)
Original requirements developed with user group input. Continue to receive input from the states.	Planned		Authorized users will have access to more data, including multiple estimates/methods for individual parameters. Public site will include only "best" estimate.	(in dev)

EIS Profile

Core Purpose	Establish comprehensive emissions estimates.
Value Added to Datasets or via Interfaces	Many estimates require complex (and obscure to secondary users) integration and calculation procedures based on reporting process and data sources, as well as high value reference value information.
Factors Influencing Sustainability of the System	Core EPA regulatory support system. Priority, schedule, and funding for enhanced access (services or applications) not yet determined. Once established, will represent the EPA authoritative source. Key issue is assuring durable performance data access to this resource.
Owners	EPA designed and operated system.
QA/QC	Documented QA/QC process on inbound emissions submissions and on estimation process.
Consequences of System Unavailability	System not yet in wide external use.

Findings and Recommendations

Terminology: EIS is a new system built to replace the old NEI system for NEI data. We will use the term EIS for the system and NEI for the data.

Development of EIS has focused primarily on support of the EPA regulatory mission. These functions include receipt, QA/QC, and integration of reported emissions estimates, as well as development and reposting of emissions factors and calculation methods. Due to the scale, complexity and implementation schedule of the EIS, the broader consideration of secondary users and integration with other internal EPA systems have received less attention to date.

A key feature of EIS is the heuristic and workflow used to compile the consolidated national estimates from the submitted data, and supplementary collected data and research. Both the estimation process itself, as well as the source and lineage of the supplementary data and reference factors are probably opaque to most secondary users. EIS is much more than just the consolidation of reported emissions.

NEI is the *national* consolidated estimate. By design, it will never be as high-resolution as estimates conducted for local areas or special purposes. Per Recommendation 2.2, we identify that an over-arching community objective is to produce the highest value, collective AQ products possible. For this data area, this means an ongoing process to consolidate from local, regional, special purpose, and other sources the best possible estimated emissions surface. This is not the purpose of the EIS system, but it is an objective the new system is well positioned to support[9].

The new EIS system provides enhanced support for emissions estimates including the capability to support, store, and assess multiple calculation methods for the same parameter space. Authorized users will be able to compare and select from among them. This capability should improve the estimation process, and provide knowledgeable users with additional options for these estimates.

Many users just want good emissions estimate data for areas of interest. Increasingly users will want these estimates as gridded products at various resolutions. For the broader AQ community, as well as for EPA internal secondary users, the overriding issue with EIS will be to ensure that they have good access to this unique data store and or its derived products. We recommend that:

- EPA should ensure a data service interface, as well as flat file dumps are available to enable/support automated or manual integration and republishing of this data.

- Access mode should, as much as possible, support users in accessing/developing the best possible estimates, for their areas of interest. Specifically:
    - Allow flexible queries for geographic areas of interest, and partitioning of emissions estimates for finer resolutions.
    - Provide access to intermediate data products used in the estimation process.
    - Support supplementation/integration with local/higher-resolution emissions estimation processes and the feedback of those products into a better national estimate.

- Access mode should expose as much as possible of the new EIS functionality and the many person-years of experience working with this data to provide data sets that are as rich as possible.

As recommended for the AQS DataMart, EPA should consider a "community informed" and "community sourced" solution for this access. EPA should consider hosting a basic set of query services outside the EPA firewall to allow external parties to develop data access applications and new web service products using EIS data. This should be developed in conjunction with an overall OAQPS data access approach, and generally should implement the coordinated services articulated under Recommendation 2.1.

Recommendations Summary

- Ensure a data service interface, as well as flat file dumps are available to enable/support automated or manual integration and republishing of this data.
- Access mode should, as much as possible, support users in accessing/developing the best possible estimates, for their areas of interest.

System-Specific Findings: RSIG

RSIG Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Varies	◒	●	●	●	○
	The current design includes the ability to provide 3-D CMAQ data. The access to this data is current limited to users behind the EPA firewall. Data from the now-discontinued EPA UVNET network is available via RSIG. The remaining data is secondary storage.	Designed to enable researchers to access a variety of distributed environmental datasets in one location. Combines remote sensing data, CMAQ data, and surface monitors including AIRNow Tech, DataFed, and AQS DataMart	Subsets large amounts of satellite data in time and space into easily accessible chunks. The satellite data can be aggregated with surface monitoring data over the same space and time, providing the novice user the ability to access these data sets in several simple steps. This system also helps connect end users with scientists.	Uses the Applet (RSIG2D). Java sub-setter publishes WCS to JAVA client running on user desktop. Open Geospatial Consortium (OGC) Web Coverage Services (WCS) and Web Mapping Services (WMS) Web-server interfaces, however it is unclear if other systems make use of these interfaces.	RSIG has a strong management chain within EPA that has provided continuity and consistent resources. However, no formal group governance of stakeholders from inside and outside the agency exists.

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
○	◒	◒	◒	●
The RSIG developers maintain a list of proposed upgrades and enhancements, based upon suggestions from all interested stakeholders	Not designed for general public, however access is available through the Environmental Science Connector (ESC)	Access to the browser-based Java applet is available only via the EPA Portal	As RSIG is currently implemented, external users are prevented from downloading CMAQ data. This restriction is not imposed by firewall or other technology, but by the owners of the CMAQ data who want to limit its distribution due to the fact that the data is of research quality. All other data sources are freely available to all audiences.	Includes ability to overlay remote sensing and other data sources into single visualization. Reduces download burden, simplifies data analysis.

RSIG Profile

Core Purpose	System designed an integrated interface for researchers to selected CMAQ output, high resolution satellite data, and corresponding ground-level measurements.
Value Added to Datasets or via Interfaces	RSIG does not perform any data estimation on remote sensing data sources. Key value-added is the access to higher resolution data, integration (via a common interface) of this data with other sources, and the re-gridding of satellite data onto CMAQ coordinates. Value is in the ability to overlay previously incompatible data sources into a single visualization. Sometimes data is reordered to match model storage and SI units, and consistent "missing value" and aggregated across many source files into a single stream. This is then made available in a variety of common formats. These include NetCDF-COARDS and optionally re-gridding to CMAQ NetCDF-IOAPI formats. Optional striding/sparsing and on-the-fly compression/decompression is employed during visualization to achieve quicker retrieval.
Factors Influencing Sustainability of the System	Funded via discretionary EPA/ORD allocation for which other projects compete. Small user base, mostly in EPA ORD and a few external partners. Not a production system.
Owners	EPA ORD and OEI designed and operated system.
QA/QC	Checks data against the source data, validates the integrity of transmitted data.
Consequences of System Unavailability	Users would have to revert to GIOVANNI's lower-resolution versions of that data. DataFed would not have access to RSIG's access point to many data sources. Data unique to RSIG, including historical UVNET and current 3D CMAQ model runs, would be unavailable.

Findings and Recommendations

EPA's Office of Research and Development (ORD) established RSIG as one of the efforts undertaken by EPA to support the Global Earth Observation System of Systems (GEOSS), to support EPA's researchers by providing access to a variety of hard to obtain data, such as satellite data, and also in support of several of the EPA-GEO Advance Monitoring Initiatives.[10] The RSIG system provides access to higher resolution satellite data than is otherwise available from the systems we examined.[11] This data is extremely useful, and its provision and use also illustrate both scientific/policy and technical challenges:

› The state of science for the estimation of AQ parameters from remote sensing data varies widely across parameters, and across spatial/temporal ranges. Some of these estimation techniques are well established and well characterized; others are still very much in development and are still actively debated by the scientific community. As pointed out in Recommendation 2.2, improving these estimation techniques and improving their application to new domains are critical steps to creating societal benefit. RSIG was designed to better link researchers exploring these areas with the underlying data. However providing "easy" access to what some might still consider "research"-grade estimated data can cause concern. Conclusions drawn using this data may be faulty, leading to poor decision-making or to undo concern about conditions that do not actually exist.

› As the resolution of data increases, so does the data volume and complexity of the routines needed to process the information. As noted above, RSIG works by "sub-setting" sources' satellite data sets into smaller chunks accessible by the user at any particular moment. The sub-setter application then uses WCS to provide a granular interface to this data.

› RSIG currently can provide direct access to the full 3-demisional CMAQ output. The ability to directly access the CMAQ data is currently being limited to access behind the EPA firewall.

RSIG has evolved over time to meet the needs of a specific group of researchers and practitioners working with ORD. To do this, the RSIG team has worked closely with end users and primary data providers, including many face to face conversations to determine which datasets can be combined to produce accurate visualizations and how these visualizations should be constructed. Though NASA continues to offer traditional ftp-whole/multi-file downloading for the remote sensing data contained in RSIG, the RSIG-developed publicly-accessible WCS applications installed at NASA provide user-application-oriented services, potentially available to both the RSIG applet and other web-based applications. Recommendation 2.1 reflects the importance of using end user feedback to assist system development and architecture, and the model that RSIG embodies is one approach to doing this.

Architecturally, RSIG consists primarily of a publicly-accessible client-downloaded applet that issues WCS and WMS requests to a publicly-accessible servlet which then streams back either data subsets or images. Behind the scenes, the RSIG server application issues WCS requests for data subsets to various internal and external WCS applications such as DataFed's and those developed by RSIG and installed at NASA. Some of these external sites run RSIG-developed sub-setter programs that perform aggregation, sub-setting, reformatting, and optional re-gridding. For WMS requests, the gathered data is streamed to an internal EPA computer and rendered into images which are then streamed back to the client for display.

RSIG's entrepreneurial success at getting data partners to install and run its sub-setter web service tool demonstrates that there is a compelling business case for this approach to expanding data access. While the specific methods and WCS extensions implemented in the sub-setter are unique to RSIG, they fully demonstrate the SoS approach. A drawback to this approach is that it requires partners to install and run the software locally, *in addition* to whatever other software they are running to source OTHER web services. With improved coordination, partners should also have the option of replicating the RSIG functions using their own web services infrastructure—the test case for this will be AirNow; they intend to mount a WCS in the near term which should provide RSIG and other users what they need. Ideally both options should be available.

The architecture, operation, and opportunistic nature of RSIG's approach reflect its nature as a research-oriented tool. It gets the job done, and because it is managed by a small group it can be adapted quickly as user feedback is received. As long as there is an understanding of the inherent uncertainties, there is no reason that RSIG functionality, particularly its WCS's, should not be more widely used. In the longer term, the current implementation may not meet the needs of expanding data sources and user needs.

OAQPS should evaluate if GIOVANNI functionality and RSIG's are converging, and identify which are the best approach to delivering these data to external applications

We recommend that EPA focus on supporting/expanding the end-to-end development paradigm while increasing the capabilities of the system.

Recommendations Summary

› EPA should focus on supporting/expanding the end-to-end development model embodied by RSIG while increasing the capabilities of the system. Gathering end user feedback directly and building system components based on that feedback is a salient recommendation and one approach to implementing Recommendation 2.1.

System-Specific Findings: VIEWS

VIEWS Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Varies	◒	●	●	○	●
	Regional Haze Rule IMPROVE data This data also in AQS	Comprehensive database of air quality data from over 24 monitoring networks. Primary purpose is to provide easy online access to relevant air quality data and tools for planners, researchers, and the air quality community in general.	Enables users to explore, merge, and analyze datasets of widely-varying origin in a consistent, unified manner with a common set of tools and web services.	Under development	VIEWS steering committee and the WRAP make recommendations and requests for data, improvements, and future directions.

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
●	●	○	●	●
Architectural and technical decisions have historically been left to VIEWS development team. The Website provides a comment and feedback mechanism for users.	Public access provided, encouraged to register		Certain users have additional privileges which allow them to add/modify web content. Agreements with various partners allowing direct database connection and data transfer. Certain resources can also be accessed via web services.	Datasets can be visualized and analyzed with the same set of online tools

VIEWS Profile

Core Purpose	Data collector which provides online access to relevant air quality data and tools with aggregation/analysis and display/presentation.
Value Added to Datasets or via Interfaces	Datasets are visualized/analyzed (sips, high grade graphics, emissions, plotting, integration tools) with the same set of online tools. Source data is altered into common schema.
Factors Influencing Sustainability of the System	Funded via discretionary but historically durable allocations from RPOs and others. Broad user base which is represented via RPOs.
Owners	EPA designed and operated system.
QA/QC	VIEWS takes responsibility for the quality of the IMPROVE datasets but for not for other, externally-originating datasets.
Consequences of System Unavailability	Major impact to AQ planners; many use VIEWS as primary access to IMPROVE network data and tools

Findings and Recommendations

VIEWS is a integrated set of applications targeted at Air Quality planners. The VIEWS application is in a regular cycle of use, feedback and revision. As important as these applications are to their community, we submit that this coupling to this community of interest is the system's key feature for the broader AQ community. The application provides a technical and organizational framework via which other services could be delivered. The ability to provide this

community with model products (like the trajectory back casts) from external services, and the exploration of using it to provide remote sensing and model are a good example of this.

VIEWS applications are tightly coupled to the underlying database and metadata model—all data is in effect "wrapped," or normalized, to this structure. VIEW's tight coupling of the data back end with the user interface provides great benefits, and means that new parameters (of the same general type) can be added easily, but that addition of heterogeneous data will be more difficult. The discussions now underway will clarify how/where this is significant.

VIEWS has some of the more complete analysis and visualization tools of the systems we examined. We considered the scenario where VIEWS was expanded to act as the AQ community's general purpose AQ analysis and visualization tool, but do not recommend this approach for the following reasons.

- It is not exactly what the architecture is designed to support.
- It may not be what the COI has an interest in supporting.
- It risks diluting the value of VIEWS to the existing COI.

Recommendations Summary

We identify the following steps VIEWS could take to increment to the preferred future (some of these are already underway):

- Identify, advocate, participate in the design of, and consume inbound web services from its current and future suppliers.

- Solicit input on the design of outbound services, for all significant internal data products produced within the VIEWS value chain. In many cases this would simply require a re-binding of existing web interface to a web services interface. (Some of this is underway now).

- Use the VIEWS user community to provide a consolidated articulate statement of users' needs and priorities to the AQ Community.

- Contribute VIEWS experience with administrative issues like data cataloging and description as the AQ community co-designs a joint approach.

System-Specific Findings: DataFed

DataFed Attribute Table

Data Update/ Generation Frequency	Unique/ Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Varies	○	●	◒	●	○
	Secondary storage only. Mediates data flow between providers and users, does not store data.	Provides access to over 100+ distributed, air-quality relevant datasets. A primary purpose to be a clearing house of data. Carries data from AQS DataMart, CASTNET, AIRNow Tech, GIOANNI, and RSIG	Basic common interface provided via a common wrapper to all services. Web services-based software which mediates between data providers and users. Users can ask clean, simple queries.	The SOA allows users to build web-applications by connecting the web service components. The data access includes protocols WCS, WFS and WMS OGC.	

Formal Feature Development Process Including Community Input	**Public Access to System**	**Behind EPA Firewall**	**Different User Access Levels**	**Integrated Interface Used for AVR**
○	●	○	○	◒
	Public access provided, no registration required			Provides standardized access and tools for data analysis and presentation. Generic web-tools created include analogs for data discovery, browsers for spatial-temporal exploration, multi-view consoles, animators, and multi-layer overlays.

DataFed Profile

Core Purpose	Began as a convenient clearinghouse for a community of interest. Republishes data due to low overhead of adding new datasets, and allows users to access data from multiple sources through a uniform, standardized interface providing access and tools for data analysis and presentation.
Value Added to Datasets or via Interfaces	Service oriented architecture. Processing and visualization services, providing tools for data exploration and analysis. Has a good metadata model, but lacks resources to develop it.
Factors Influencing Sustainability of the System	Connection to EPA sources. Many services do not work reliably, some use outdated data sources. Dependent on episodic and indirect funding to host. Established as an operational prototype to demonstrate technology.
Owners	Independently owned, hosted at Washington University
QA/QC	No quality assurance or integrity checks, uses data as-is.
Consequences of System Unavailability	Loss of standardized processing and visualization services to analyst community of interest.

Findings and Recommendations

DataFed is a powerful technology and architectural demonstration of SOA in action. It demonstrates the:

› Ability of a SOA resource to provide significant value to the community just by helping them locate and bind to existing data resources; none of the data available on DataFed is available only via DataFed, but for many sources, it is the most popular access mechanism.

- Ability of a third-party mediating service to provide flexible access almost any data resource that almost any kind of reliable connectivity is provided for (files, REST or OGC WS).
    - This mediation is achieved through by applying a standard metadata "wrapper" around all services so they can managed from a single navigation/location service (see below).
- Ability to provide basic analytic services like reformatting and re-gridding over a standard interface.
- Utility of having a single, service-based locator service that is extensible with all manner of metadata.
- Ability to map and graph almost any data from these sources.
- Utility (and drawbacks) of using social collaboration for COI information exchange and information/metadata development.

Examples of DataFed's mediation and processing functions include:

- **Hosting of AQS web services**: DataFed periodically retrieves AQS flat files, loads them into a local data store and republishes them as web services.

- **AirNow Re-distribution**: AirNow ASCI OBS files are compiled into binary cube (for fast access) republished as WCS then a WMS for ad-hoc query. This data is used, for example, on GIOVANNNI.

In addition to these public interfaces, the DataFed platform also supports collection of specialized modeling and visualization research applications used by staff and collaborators, these are the "production" applications supported by the platform.

As indicated above, we see DataFed as a powerful validation and illustration of the overall architecture of the preferred future SoS. Nearly every function envisioned in a near term preferred future AQ SoS is operational in some form on DataFed its services are used by many of the other systems assessed, and nearly every system assessed is available via the DataFed interface.

DataFed has demonstrated all this but its public interface is not a production system. As staff readily admit, much of the public resources of DataFed have been stitched together via work on various projects, and have not received dedicated funding and staffing outside those projects. Some of the interfaces on DataFed don't work at any given point, and for some of them, the versioning and documentation of lineage is done on an ad hoc basis. Many of the feeds to DataFed are performed semi-manually and DataFed does not perform QA/QC on the refresh process beyond that provided by the transfer process itself.

We point these issues out because we believe some members of the AQ community are interpreting glitches or deficiencies in the current DataFed as deficiencies with the technology or worse, and incitement of the whole concept of SOA. This is counterproductive. If potential partners hear that the future is DataFed and they don't happen to like the way DataFed looks or works that day, they may say, "I don't like your future." This would be a waste and a distraction.

Recommendations Summary

- EPA should consider DataFed as an operational test bed for demonstrating rapid delivery of data and analysis services.
    - EPA should consider providing some targeted support for development of a dedicated application. This could include soliciting user requirements and prototyping a general purpose data browser and data download tool. AQS staff ID that many users still seek simple exploration and DTF file download; these could be easily mediated by DataFed via the existing AQS DataMart EN service.
- DataFed demonstrates the value and appetite for data discovery and exploration services. These could be adapted from DataFed and promoted/re-deployed to a "production" role for the SoS and operated, officially, on behalf of the AQ Community SoS. The underlying model is sufficiently compliant with evolving OGC standards—it already includes spatial-temporal data concepts and can easily be extended to include the method/parameter based search (e.g., find all data sources for a given consituent). This effort could be linked to and feed from existing metadata repositories (where they exit) and produce demonstrable value with a very limited input of Agency resources and staff time.
- Many ongoing projects are considering or are using parts of DataFed; EPA should try to coordinate with those partners, perhaps with some targeted resource supplements to elevate the new products developed into the first generation production environment for the AQ SoS.

System-Specific Findings: GIOVANNI

GIOVANNI Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
Varies	○	●	●	●	●
	Merges 3-4 datasets from different sensors. Primary source estimated. Primary storage of OMI, MLS, HIRDLS, AIRS, TRMM, TOMS.	Designed to provide a user interface to preexisting datasets	Can be used by anyone through a Web browser	Provides WCS and WMS access; will try to provide GIOVANNI as Open Dap web service. CGI scripts written in Perl and in GrADS	NASA Operated

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
●	●	○	○→◒	●
User groups for each instance to provide guidance and feedback	Public access, does not require registration	Located at NASA	There is a category served through a different port, mostly to support science teams who don't want the previous data to be available to the public	

GIOVANNI Profile

Core Purpose	Created to provide re-gridded satellite data combining multiple datasets.
Value Added to Datasets or via Interfaces	Major source of Level 2 satellite data at NASA. Value add of the system is re-gridding different datasets to be comparable. Users can explore, inter-compare, and analyze large amounts of Earth Science Remote Sensing Data using a web browser with no download required.
Factors Influencing Sustainability of the System	Core part of NASA data access architecture. GIOVANNI is regarded as having the most sophisticated UI of any systems we examined. Other systems are beginning to provide similar data from NASA DAC.
Owners	NASA Goddard.
QA/QC	Quality assurance not provided.
Consequences of System Unavailability	Reduced access to Earth Science Remote Sensing Data.

Findings and Recommendations

The principal value-added of the GIOVANNI[12] is the interface and data processing which happens before results are displayed on the screen. GIOVANNI does very complex re-gridding of data from a wide range of satellite sources. The data is able to be queried on a wide range of parameters that are adjusted depending on the data source, improving usability of the

system. Results are presented through a map interface, or downloadable in CDF, NetCDF, KML and other common formats for easy integration and re-publishing.

For most systems in common GIOVANNI courser resolution (spatial and temporal) data than does RSIG, which also provides remote sensing data from NASA. However, GIOVANNI is planning to implement access to finer-grained data in the near future. This is an opportunity to work together and provide value to customers from both tools, or combine elements of the tools.

GIOVANNI implements and consumes a wide array of Web Coverage Services and Web Map Services. Among the systems we examined, DataFed is one of the largest consumers of GIOVANNI data, and GIOVANNI pulls AirNow PM2.5 re-gridded data from DataFed.

Recommendations Summary

- We have no recommendations for GIOVANNI beyond the following general observations:
  - RSIG should confer with GIOVANNI to understand convergence of systems.
  - GIOVANNI staff should participate with the AQ community in consideration of future protocols beyond WCS 1.0 to make parameter-driven queries more functional.

System Specific Findings: HEI

HEI Attribute Table

Data Update/ Generation Frequency	Unique/Primary Data Provider for that Assemblage of Data	Created as a 'One Stop' for Data From Multiple Systems	System Includes "Novice" User Friendly Integrated User Interface	System Includes Outbound Web Services (used by external customer)	Formal Group Governance/ Process
annual	○	●	●	○	◒
	Secondary data only from AQS DataMart, Census Bureau, and NCDC	Purpose is as a 'One Stop' for health COI; acts as clearinghouse	Access through interface, includes site browser, list building, database queries, and users' guides		Decisions made by HEI

Formal Feature Development Process Including Community Input	Public Access to System	Behind EPA Firewall	Different User Access Levels	Integrated Interface Used for AVR
●	●	○	○	○
Research Committee provides comments and suggestions – but whatever the clients say, they implement	Access to registered health researchers	Hosted by AER for HEI		

HEI Profile

Core Purpose	Establish a "one-stop" data source for health COI.
Value Added to Datasets or via Interfaces	Unique statistical summary values were generated for all data in HEI to facilitate cross data-set searches. The interface allowed for dynamic queries and retrieval of data in CDF format.
Factors Influencing Sustainability of the System	Dependent on funding from multiple sources. Funds were not extended past 2008. Convenience system for users.
Owners	Health Effects Institute owns the system. Atmospheric and Environmental Research (AER), a consultancy, hosts and operates the actual HEI database.
QA/QC	Data entering the system was tested for completeness and junk values. No comparison was run against the source system.
Consequences of System Unavailability	System has been de-funded, so we will find out. Users will probably begin to call EPA and other sources for the AQ data they previously got from HEI. As discussed, this represents an opportunity to demonstrate a SOA approach to serving this need.

Findings and Recommendations

HEI has been successful as a simple, easy to use access point for air quality information of particular interest to the health research COI. It was designed and built to create a "one-stop" system because there was no simple way to access this data through other interfaces. The data aggregate by HEI includes AQS data accessed through a SQLNet connection to the AQS DataMart, and meteorology data pulled from NCDC.

HEI is accessed through a web interface that allows for basic searching by pollutant over a span of time. The results are available in multiple file formats, but CDF is the most popular. XML was considered as an output, but was seen as too complex and with too large an overhead for the

target community. Common tools in this COI can natively import CDF, while XML would need to be de-normalized to be readable.

The key value adding component for this system was simplicity of data and interface. The data requires only basic transformations and interface to be useful. Data was used for analysis and research, so HEI provided a low overhead interface for users.

Surprisingly, HEI was used by a wider range of users than anticipated, including EPA and state staff. This indicates that there is still an appetite for easy access to data downloads from the AQS DataMart. Development of an HEI-like interface for data downloading using WCS should be investigated (an idea also discussed at the Data Summit).

HEI reinforces the general finding of this report that specific COIs benefit from custom interfaces—even when interfaces provide only small customizations, those customizations can be critical for that COI. The architecture of HEI (as a manually-updated data warehouse) needlessly replicated the functionality of other systems, given the low amount of value-added processing that was done to the data. The "one-stop" nature of the system could just as easily be recreated with dynamic queries to primary source systems (like the AQS DataMart). The availability of a simple interface and access to simple data formats would meet the needs of a specific COI and require relatively little development work.

Recommendations Summary

› Per Recommendation 2.2 there are several products of interest to the health community Given the historical importance of HEI to a specific COI, we recommend that EPA explore ways of providing the current functionality and value-added of HEI processing through an existing system like the AQS DataMart, alongside links to the best known estimated surfaces.

APPENDIX C: SYSTEM SPECIFIC RECOMMENDATIONS SUMMARY

System	Recommendation
AIRNow	• AIRNow should establish a WCS service for use by RSIG and others. • AQS and AIRNow should coordinate on a simple automated refresh procedure.
AirQuest	• AirQuest and its use should be re-examined in the larger context of OAQPS systems. • Broader planning effort that initiated AirQuest should be renewed.
AQS/AQS DataMart	• The AQS DataMart is the main interface for Air Quality information and should remain so. • A simple, user-friendly interface should be developed to access simple, low-/medium-end access to this data. • EPA should consider a "community informed" and "community sourced" solution for access; possibly through a basic set of query services outside the EPA firewall to allow external parties to develop data access applications and new web service products using AQS data. • Enrich AQS metadata by improving accuracy and volume of self-reported metadata and make accessible through a publishing mechanism a service to all DataMart users.
CASTNET	• EPA should work with CAMD to scope the potential for integrating CASTNET data into the DataMart. • Make CASTNET data more easily available for republishing. • Include CASTNET in future planning for improved access to modeled data products.
EIS	• EPA should ensure that a data service interface and flat file dumps are available to enable/support automated or manual integration and republishing to EIS data. • Access should support users in accessing/developing the best possible estimates for their area of interest. • Access mode should expose as much as possible of the new EIS functionality. • EPA should consider a "community informed" and "community sourced" solution for access; possibly through a basic set of query services provided outside the EPA firewall to allow external parties to develop data access applications and new web service products using EIS data.
RSIG	• OAQPS should consider if GIOVANNI functionality and RSIG's are converging for determination of which is the best approach to delivering these data to which future external applications. • RSIG functionality, particularly its WCS's, should be more widely used. • EPA should focus on supporting/expanding the end-to-end development paradigm while increasing the capabilities of the system.
VIEWS	• Solicit input on the design of outbound services for all significant internal data products produced within the VIEWS value chain. • Identify, advocate, participate in the design of, and consume inbound web services from current and future suppliers. • Use the VIEWS user community to provide a consolidated, articulate statement of users' needs and priorities to the AQ community. • Contribute VIEWS experience with administrative issues like data cataloguing and description as the AQ community co-designs a joint approach.
DataFed	• EPA should consider DataFed as an operational test bed for demonstrating rapid delivery of data and analysis services. EPA should consider providing some targeted support for development of a dedicated application. • Discovery and exploration services should be adapted from DataFed and promoted/re-deployed to

System	Recommendation
	a production environment and operated officially on behalf of the AQ community SoS. • EPA should coordinate with partners of DataFed, perhaps with some targeted resource supplements to elevate the new products developed into the first generation production environment for the AQ SoS.
GIOVANNI	• RSIG should confer with GIOVANNI to understand convergence of systems. • GIOVANNI staff should participate with the AQ community in consideration of future protocols beyond WCS 1.0 to make parameter driven queries more functional.
HEI	• EPA should explore ways of providing the current functionality and value added of HEI processing through an existing system like the AQS DataMart.